Personalization of Web-based Interfaces for Humans and Agents

Including Aspects of Privacy and Applied to E-Government Portals

Michael Sonntag, Jörg R. Mühlbacher, Susanne Reisinger

{sonntag, muehlbacher, sreisinger}@fim.uni-linz.ac.at

***Abstract**:* An important part of E-Government is bringing the administration closer to citizens. A way to do this are webportals, where information and (partly and in a small scale now, increasingly in the future) online transactions are possible. To increase the utility of these portals for citizens, personalization can be used for presenting more tightly focused information. As customers of such services can be both humans and agents, In this paper we present the methods possible for identifying and adapting the content to them afterwards. Privacy issues of personalization are also looked into, according to the EU directive. Necessary aspects of personalization for E-Government portals are identified and applied to the methods presented: For which area which type of identification and adaptation of content should be used.

Keywords: Personalization, Webportals, Agents, E-Government, Privacy

Introduction

An important part of E-Government is granting citizens and businesses access to information on administration, proceedings (those they are party to, as well as generally), and probably even transactions, online and not only during office hours but 24 hours a day ([Wimmer/Krenner 2001]). This is usually done through the WWW, as it is a universal means of presentation and widely available and known by the largest percentage of the population.

We consider mainly the conventional situation, that a provider offers and transmits information on request after a successful login of a user. But also the ability of pushing information ([Bonifati et al 2001]) to the (previously identified) clients, based or triggered by some events at the servers site, is faced with the aim of personalizing web based interfaces. We also use the term „personalization“ if the user actually is an autonomous agent. In this case personalization refers in particular to a specific adaptation to the capabilities of the agent in question. Therefore whenever we speak of a “user” we include addressing an agent as well, whenever it is meaningful.

Referring to agents we use the broad characterization of [Wooldridge 1997]:

> “An agent is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its designed objectives.”

It is evident, however, that some of the methods, as they are described below, make sense only if the user “behind” (or in front of the webportal) actually is a human being, connected by some client interface.

At least transactions of all possible interactions require identifying the user. If the user is already identified (or if otherwise recognized as the same visitor as some time before), the information presented can also be adapted to his needs, further improving the quality of the service. Adapting the website can be done in different ways:

- Adapting the presentation: E. g. for people with disabilities a text-only presentation, larger fonts, special colors, etc. could be provided
- Adapting the data: For minorities or foreigners, providing information in different languages is an important aspect.
- Adapting the information: Content can be selected according to stated or derived (from various sources) areas of interest.

- Assisting the user: Giving hints on where information, which might be of interest can be found, explanations on what to type in a certain field of a form (perhaps in different styles of language for inexperienced and professional users, examples for applications), etc.
- Access through agents ([Jennings/Wooldridge 1998]): Beside access by humans, portals should also be accessible through autonomous software (commonly called agents; [Mühlbacher et al. 2001]). In this case personalization is even more important, as interpreting general information as relevant is more complicated and assistance by the portal can greatly ease this.

In this article we will therefore take a look at some issues of personalization of webpages with a special view on E-Government portals. The aspects covered include how to recognize or identify the user, adapting the content to individual users (selecting information/data from a larger store, presentation is not covered herein) and legal issues of personalization (privacy).

At the end we will identify some important aspects of personalization which must be heeded for E-Government portals and present some recommendations how to identify users and which methods of adapting the content might be most appropriate for certain elements of E-Government portals.

Recognizing Users

To be able to adapt the content to individual users it is necessary to first recognize the user as a certain singular individual. This might be done in two different ways: Recognizing the user as the same person as before (previous visits; relative identification), or recognizing the user as a certain person with its data (absolute identification). The other alternative, always creating a "new" personalization works only with very simple types of it, as always entering or observing the data anew will allow only little information to be deduced, resulting in poor quality and questionable added value to the visitor. This added value is a very important part, which may never be neglected ([Wenger/Probst 1998], [Sonntag/Reisinger 2001]), because if there is no advantage in using the portal, citizens will continue using conventional means of doing interaction with administration.

In the following an additional distinction always has to be remembered. In many cases not the user is identified but rather the computer the user sits before (which might be a difference):

1. There are users with a number of computers (company computer, home computer, laptop, PDA, …) which could be potentially identified as different persons. This might be adequate because of different areas of interest (professional and private), but in most cases this is not desired. In E-Commerce these are the most "wanted" customers and in E-Government those, for which the service is most important and potentially useful.
2. There are computers used by a defined community, e. g. home computers, which are available to every member of the family. And there are situations where an anonymous number of persons have access to, as it happens to be at public terminals. In the latter case persons should be treated as different users for personalization, and in terms of and with respect to privacy they must be treated as such.
3. In the case of an autonomous agent, the recognizing procedure seems to be easy, based e. g. on an unique identifier assigned to the agent when it was registered. However the tasks becomes recursive as soon as an agent is being used on behalf of persons.

Username/Password (HTTP, Forms)

A very simple and common form of identification is, that every user selects or receives a username and must provide the corresponding password on login. This can be done either through a form on a webpage or directly through the HTTP protocol. The advantage of this

method is that the user is uniquely identified as a certain person (not the computer). Problematic is, that HTTP is a stateless protocol, so this method alone allows personalizing only a single page: the page presented in result to the form (and subsequent pages, as long as there is a "chain of forms", unbroken by direct links; when using HTTP authentication, the username/password is stored and passed on each request). Therefore this method alone is not sufficient and must be combined with others.

However, a password is easily forgotten and can be passed on to others. Also problematic is that most users will be using many services. This results in either many different passwords, which will lead to a written "help-sheet", which is bad from a security viewpoint, or the same password for all services, which is also no good idea. One security leak leads to a large security hole. This can happen easily if only a single service uses unencrypted transport. Storing all passwords in a single place or using a common method of login (e. g. Microsoft Passport) has also many problems. Users are also prone to forget passwords, so additional hints are necessary:

1. The password is sent to an E-Mail address entered at the registration. Only the real user has access to this mailbox. This is a reasonably safe assumption, because of the huge number of E-Mails in the Internet. However, it is sent unencrypted and could be filtered out in transit or stolen from the mailbox.
2. Additional hints or questions are shown. The user can enter a hint on registration or answer some (possibly predefined) questions about him. Both versions are dangerous: A hint can be meaningful for many persons, not only the real user, and must be shown publicly. Also, answers to questions might be known on a broader scale (common: maiden name of mother, name of pet, …) and are additional personal data, which the provider usually has no real need to know (privacy).

Security in this area is usually rather low, except on E-Commerce websites, which also use encrypted transfer because of users providing creditcard numbers. In all other cases the username and password is sent unencrypted. When using HTTP this is even worse, as it is sent on each request (webpage, image, …; unless a persistent connection is used). In HTTP version 1.1 encrypted transfer is possible but depends on the capabilities of the participating systems. This method, however, is only seldom used. The identification is used mostly for allowing access according to file-system permissions. This approach has the property, that the identification is only stored for the lifetime of the browser. The advantage is that different persons on the same computer are identified as different users (if the browser is restarted; a problem possibly on public terminals), but that a user always has to reenter the information on first access. Also, HTTP does not allow custom pages for entering the user data: The user must enter the data before anything is displayed, so he knows only from the URL used where this data is sent.

Cookies

Cookies ([CookieSpec], [Kristoll/Montulli 1997]) are short text files stored on the user's computer (therefore permission for this storage is required) or in the local storage space an agent may provide for information exchange. Any information can be stored in there, for example an encrypted (or unencrypted; name, …) identification number, as long as it is rather small (limitation to 4kB per cookie). The use of cookies usually is transparent to the user: He is (if not configured otherwise) neither informed when cookies are stored nor read from his computer by the webserver. Disadvantages of cookies are:

- Because of misuse, some users do not allow cookies on their computer. In combination with a lot of servers using them, manual (dis-)allowal is rare.

- In the standard for cookies no maximum size is defined, but most implementations use the minimum size (4 kB, 20/server) as the maximum. The data for personalization can therefore not be stored in the cookie, only some identification mark.
- Cookies possess an expiry date (set by the provider) and can moreover be deleted at any time (e. g. if disk space or memory is needed). The server must therefore cope with disappearing cookies, even during a session. Because of this the website must provide at least one alternative form of identification.
- With this method, not the user but the computer is identified (copying the cookie files to another computer is possible but never used). For shared computers a possibility should exist to explicitly remove a cookie (e. g. some form of logout), otherwise subsequent users are seen as the first one by the provider.

In spite of these problems, cookies are still the most often used and possibly a reasonably good solution if they are combined with another method, as no interaction of the user is required and secure and reliable identification is possible including exchanging personalizing information.

If agents access webportals directly, they must also be prepared to deal with Cookies (see for example the Agentsystem POND [POND], where this is included and applied in practical agents). As they can easily implement filters (which cookies to accept, when to delete them, to which server to send them, ..), there are advantages for privacy. So can e. g. external cookies (which are a strong hint for advertisements) be filtered out automatically.

URL-Encoding

A number identifying the user can also be encoded in an URL (similar to submitting the content of a form using the GET method). A result of this is that the URL is rather long (the number is additionally encrypted in some way usually) and almost impossible to remember. Agents possess an advantage here as this is no problem for them. Using different computers is therefore difficult and personalization is useable only when the webpage is accessed through a bookmark (nobody will want to enter such an URL by hand). An advantage of this method is that absolutely no special software or technology is needed, so this method is ideally suited for fallback if other methods do not work (in combination with another method for initial identification). However, only a single page can be personalized or the software must modify all internal hyperlinks on the webpage to also include the encoded identification mark. If there is some kind of contract through a protocol, agents could supply this additional data automatically. Also, only knowledge of the URL is tested, which is the same as the password, so unique identification of users is possible, but rather insecure. While a password can be remembered, such an URL can probably not. So everybody with access to the list of bookmarks can pose as the list's owner. To avoid this, random long numbers are used. This requires a different method of identification for initialization (e. g. passwords), while all following pages can then be personalized in this way.

Unique identifying numbers

Identifying the computer can be done through some unique identification marks, like its IP-address, the MAC number of the network card or the serial number of the CPU (in newer models). The later ones are rather difficult to detect, in contrast to the first. However the IP-address is in many cases not unique, as access providers dynamically allocate them, and possibly proxies are used, which also disallow identification in this way. Also there is no way to distinguish whether this is a static address or not, and whether it is used for a single computer or many (proxy). The serial number of the CPU cannot be read remotely, so local software would be needed (see below). Because of this, rather secure identification (which is

required in E-Government) is impossible and, anyway, only the computer would be identified. Additionally, severe privacy problems exist ([Brandl/Mayer-Schönberger 1999]).

Add-On Programs

Using add-on programs is another alternative for identifying users, which is however rather rare and not appreciated by users. This means installing a program provided by the operator of the website on the local computer, which users are hesitant to do (viruses, trojans, backdoors, etc.). The program can be either a custom browser or any other program. A disadvantage is, that this program must be installed on every computer a person uses and that the internal data of the program must be synchronized between them. This program can easily detect the user logged on and can then use any secure protocol (e. g. public key cryptography) to deliver this information to the server, exactly identifying the user. Several drawbacks of this method exist:

- An additional program is needed, which must be different for all platforms. Also versioning problems may arise (distributing new versions, coping with old ones, …).
- No general standard exists. Custom software must be developed. The software is useable only for a single provider.
- Multiple instances and synchronization is needed if multiple computers are used.
- Use in LANs can be difficult, as Proxy servers and/or firewalls are used. As custom protocols are efined, this possibly opens up new security holes.
- Users are generally reluctant to install programs of any provider unless there is already some kind of relationship with them, e. g. a contract or it is a publicly known institution (as might be the case in E-Government).
- Is only the identification sent or other data too?

If this method is included in a separate program (client-server model), no problems exist (e. g. closed systems like notaries or courts), but this is not useable for the public: A common and well-known interface must be used, which is almost always a webbrowser.

TLS / Electronic Signatures

TLS (former SSL; [Dierks/Allen 1999]) is a protocol, which can be used below HTTP to provide encryption and identification of parties. Commonly only the server is identified, but the protocol also allows mutual identification (identification of the client = user alone is not possible). However, as it is independent of HTTP, using it for personalization might be difficult because of the interface needed. Using electronic signatures is also not really a viable alternative, as each request would have to be individually signed and also the interface-problem exists. An issue in this connection is, that usually no “anonymous” certificates are available; at best a pseudonym can be used. This results not only in recognizing a user, but identifying him as a certain person with all additional information possibly included in his certificate. If the contact to the server is mediated through an agent, many of the problems disappear: An agent can create and store alternative certificates used for identification (perhaps only to a single provider) and automatically provide them according to his internal rules.

Persistent HTTP-Connection

From HTTP version 1.1 upwards, a client can use one connection for multiple requests to a server. This is implemented by leaving the connection open after serving a request. However, it can be terminated at any time from either side. Normally clients use this is only for requesting a whole page with all it’s associated content (e. g. images), but close the

connection afterwards. This is therefore not really useable for personalization as only a single item to personalize (webpage-content) is transmitted each time.

Methods of Adapting Content to Users

Several possibilities exist to adapt the content to a user once he or she is recognized. In most cases a relatively large amount of work is required for classifying the content to personalize, so it can be adapted later to the preferences/interests of users. Often a combination of at least two methods described will need to be used as each has at least one important drawback.

For all method advantages and drawbacks are presented after a short description.

Questionnaires

The most common way of adapting content is to explicitly ask the user, what he is interested in. Usually this is done by filling out a form on a webpage where interests, dislikes and so on are stated or can be entered. The actual selection of the content is rather easy then, as the category of the data need only be matched to the categories stored with the user's data.

This is the method best suited for agents as a recipient of the information. It can explicitly state the information he is interested in (similar to subscribing a service), or the areas his owner is (probably: if gained through observing his behavior) interested in. This is also an advantage for the server, as he already receives a pre-controlled and focused set instead of having to collect data through forms (or observations; see below).

Advantages

- This method is easiest to implement and requires the least amount of storage and computing time.
- Preferences can be asked for very easily at registration or through a protocol.
- Classes for the classification of the content are simple to create according to the questions.
- Rules for selecting the actual data presented are simple to create.
- Restricted prediction of other interests of users are possible: If e. g. the user likes A and B, he will probably also like C. Finding these rules, however, can be difficult.

Disadvantages

- Personalization is relatively weak, as only a limited amount of information on the user is available. Nobody will fill in 10 pages of questionnaires if he is not forced to, especially before he receives any (possible) advantage from it.
- If the user does not take actions to change his areas of interest, the data on him stays the same. When his information needs change, personalization grows less useful to him over time (and they will change).
- For useful personalization complex classification is necessary. This means a lot of work in advance before any users can utilize it. For large amounts of data this is even worse, as data must be classified consistently, even if done by a number of persons (which is difficult).
- Users don't always disclose their real interests or all of them. In this case personalization is even detrimental.

Search-Path Shortening

This method of personalization adapts the content to the user by shortening his path through the tree-like structure of a website to the leafs containing information: The more often he clicks on a link, the higher up it will be placed. This can be done either by moving it up in a list or by transporting it to a higher level in the hierarchy. Through this, the information of most interest to a user slowly moves to the top reducing his search-time for consistent requests.

Advantages

- This strategy continually improves itself and therefore automatic adaptation to changing interests of users is achieved.
- Personalization is done according to the real interests of a single person, without influences from other users or profiles.
- No classification of the data is required and every type of data (e. g. also documents, images, messages, applications) can be personalized.
- No explicit actions by the user are necessary.

Disadvantages

- Personalization takes a longer time to be effective and changes only slowly.
- Only persons with high usage profit from this, as passers-by see no results.
- No predictions are possible, what might be additionally of interest to the user.
- Users might be irritated if links important for them continually move around. This is a huge problem for agents, as finding the appropriate links is even harder for them.
- Usably only if the same or at least similar items are requested repeatedly.

Collaborative Filtering

Many users rate the content and according to this rating they are divided into separate groups using statistical methods. For personalization the prospective user must rate some selected elements and is sorted into a group according to his answers. All the interests of this group are then defined to be also his (if the user likes A and B and his group likes A, B, and C, he will probably also be interested in C). In some variations, the user is not sorted into a group, but a single user, which matches his interests best, is selected (a virtual “twin”). Allowing agents to rate information is dangerous, as another layer of possible misunderstandings and inconsistencies is introduced.

Advantages

- No classification of the data is needed and all different types of content can be included.
- The more persons use personalization and the more they rate, the better it gets for all, resulting in continuous improvement of the personalization.
- Predictions for different areas are possible, as long as they are rated by the group the user is sorted into. Also the probability of this being correct can be calculated.
- With only a few rates relatively good personalization is possible, as long as a large user-base is available.

Disadvantages

- The algorithms involved are complicated and require a lot of resources (CPU-time).
- Creating groups is the key point of this method, but cannot really be predicted. Some groups might be very good, but others might consist of very different interests.
- A large number of regular users must continually rate items for good results. Even if the personalization already works well for them, users must continue to rate items.
- Starting is difficult as there is no data available then. Possible solutions for start-up possess low quality.
- Only long-living items can be personalized in this way: E. g. daily news are of no interest any more when there are enough rates to allow personalizing them.

Observing Behavior of Users

This method starts with an unpersonalized page and then observes the users behavior: Where does he click, how long does she remain on certain pages, what words does he search for, from which websites did she come, interactions done with this website, and so on. According to predefined rules from this data the areas of interests of the user are deduced. Afterwards content is selected according to a comparison between its classification and the interests found.

Advantages

- The real interests of the user are found, according to his actions.
- Personalization starts without need for any additional action by the user.
- The content is adapted to the interests of the single user, not a group.
- Continuous improvement is possible, as long as there are applicable rules.
- The behavior of agents is strictly rational and therefore interests are easy to derive.

Disadvantages

- A lot of personal data must be collected (clicks, time, data entered in forms, …) and analyzed afterwards. This clashes with privacy issues.
- Some data can be misleading too: Keeping a specific page opened for a long time simply might refer to a “coffee break” or other work done concurrently.
- Collecting sufficient information for effective personalization is time consuming. It is more than just waste of bandwidth. It keeps the user busy waiting without presenting an immediate advantage. It is therefore usable only for “regular” visitors, which enjoy the personalized interface the next time.
- If the content is already personalized, deducing interests from actions can be rather difficult. Predictions on new areas are not possible.
- Finding the rules, which actions on which elements/classes imply which interests is difficult. Sometimes this can be only done during actual deployment.
- Content must be classified. The categories for classification might not be complete at that time, so possibly even re-evaluation is necessary.

Statistical Profiles

For this method only very few general questions are presented to the user, which might be rather different from the content to personalize. According to very extensive statistical profiles, the user is sorted into a certain category and all empirically found interests of this class are ascribed to him. With respect to agents this works only if they represent their owners directly and have been "fed" with personal information before having been sent on their way.

Advantages

- Only very few questions to the user are needed.
- A very broad spectrum of very detailed interests is associated with users.
- Predictions on completely different areas are possible.

Disadvantages

- Selecting the questions for sorting users is very hard.
- A huge amount of statistical data and much experience is needed for creating classes.
- Nobody exactly conforms to the statistical average. In some areas the profile is wrong, and these might even be the most interesting ones. Detection of this problem is impossible in advance.
- Once sorted into a class, no further changes are made unless the statistical group changes.
- No adaptation of the personalization to a single person, but more an adaptation of the user to a class.

Privacy Issues of Personalization

Personalization is typically opposed to privacy: The former tries to accumulate an ever increasing mass of data on a person, while the latter tries to keep personal data out of other collections of data. As privacy is not only a personal issue but has been put down in laws ([DSG]), including directives from the EU ([DS-RL]), providing a common lowest level of protection, these regulations must be taken in account when implementing personalization.

Classification of data

Data must be classified into four distinct groups according to the Austrian Privacy Law and the EU Directive with respect to privacy (this is different from the classification according to the content, which is an aspect of most methods of personalization):

1. Anonymous data: This data cannot be connected to a single person by anyone. It is therefore not protected at all. Examples are website usage data after deleting the logfiles. This data is not of interest for personalization.
2. Indirect personal data (§ 4 Z 1 SigL): This is personal data, which the current owner cannot ascribe to a single person through legal methods and reasonable effort. Examples are website usage data while logfiles still exist (Tracing the data back takes a lot of work and is only possible with logfiles from e. g. the users Internet provider, which is not allowed to share them with others).
3. Personal data (§ 4 Z 1 SigL): The largest class, which encompasses the "normal" data like interests of registered users. This applies even when the single person is not known by name (only e. g. E-Mail address), or not unambiguously (e. g. very common names).

4. Sensible data (§ 4 Z 2 SigL)/Special data (§ 18 Para. 2 SigL): Personal data on special issues like race, political opinion, health, as well as criminal history or creditworthiness. This is a closed list in the law. For E-Government the areas political opinion and membership of labor unions are most important. For this data there is additional protection (e g. permit required in advance).

Responsibility

Responsible for meeting the legal rules is the client on whose behalf data is used, even if someone else does the actual work ([Jahnel 2001]). Only if the sub-worker processes data against explicit prohibits by the client, or according to legal requirements, he acquires the legal status of a client and must take responsibility for these actions.

For E-Government this usually results in responsibility of the central government. However, this is a difficult area as a webportal for E-Government might include also issues, procedures, forms, etc. of regional or local government. If the portal only provides a link to pages, which are maintained by the other authority, there is no problem: The government of the linked-to site is responsible. If only general information (or forms to download, print and then filled-in) is provided on the portal, this is not a problem of privacy so no difficulties arise there (in this connection; liability for incorrect/misleading/incomplete information is a different issue altogether). When the form is filled in online and the data then passed on to other governments, these regional/local governments are the clients and therefore responsible while the central government (or who actually runs the portal) in this case assumes only the role of a subcontractor. Because of this, for a unified portal it is important to clearly mark where certain data is used and by whom, because only then the obligation of informing the person is fulfilled (and the person can then use appropriate measures if he suspects misconduct).

With regard to the data used for personalization itself (e. g. obtained through observing the user), the operator of the website is the client.

Gathering data and user's consent

Gathering data is an important part of personalization. Privacy however forbids this if not an exception is applicable. The most important ones in this case are legal requirements (the citizen is under an obligation to disclose this data) and consent. Consent is only then valid, if the person is exactly informed which data is collected, in which way this happens (e. g. from where), and what it will be used for. Misunderstandings and duress disallow consent. The area the data will be used for must be specifically stated, but it might be rather broad. A general consent for all uses is not possible.

The classification of data is important here: For "normal" data, consent can also be conclusive, while for sensible or special data it must be explicit. The latter need not be in writing, but this might be useful for proving the consent. This could be done in a portal through electronic signatures. Existing software allows signatures only for documents and E-Mails, but no integration into the WWW exists, so this is not an alternative. Presenting the user a form where he must actively change something (e. g. click on a checkbox) will be sufficient for explicit consent.

Data can also be obtained from other sources, e. g. regional governments or private companies, as long as there is a legal cause for this (for the transmission; authorization for possession and use alone is insufficient). An alternatives, if the person has given consent to this change of use (which is to be determined according to its use at the previous owner and the possible, but not to be presumed, consent of the user for passing it on).

The same also applies to cross-connecting data from different sources. Consent of the user is needed for this, otherwise each part of the site may use only the appropriate data (e. g. local

government data might be used only for assisting users when filling in forms directed to the local government, but nowhere else). This consent should always be obtained, because partitioning the data in this way will lead to many misunderstandings by users and also possibly to inconsistencies.

Passing on and changing the use of data

The same issue as applies to cross-connecting data is also valid for actually transferring it, e. g. when a form is filled in by the user and additional data is sent with it (for example confirmation of certain data like nationality or certain admissions).

Obtaining is somewhat easier here, as it can be presumed to be implicitly given for certain data. Items which are filled in are obviously contained, but also any data explicitly referenced. Not included are confirmations for data without mentioning that it will be automatically sent along and of course sensible data. In other cases, data may not be sent to any other agency (e. g. for studies, testing, or use) without first obtaining permission.

In connection with personalization it is important to mention that even simply changing the area of use of data is legally seen as a transfer. Using officially available data (e. g. field of business, amount of taxes to pay, or trade admission for companies; number of licensed pets, areas of work, social security benefits for private persons) for personalizing the portal is therefore also bound to the permission by the user. Simple registration at the portal will probably not be enough to construe consent. Also, choices which data to use (according to both the source and the content) should be possible. Only this data is then used to adapt the content of the portal to the user.

Similarly, data gathered through the website, e. g. which pages/forms the user viewed or started to fill in but then abandoned, may not be shared or used for anything other than personalization. As this would constitute a completely different use, no implicit consent can be assumed and explicit permission must be obtained for this. In terms of privacy and citizen-friendliness this should be avoided altogether, however.

Privacy and agents

If data on the user is passed on to the portal, it is obviously protected even though it might be rerouted and provided by an agent. Data on the agent itself, however, is not that clearly protected as an agent is no person (neither a natural nor a legal one). But the agents behavior is still attributed to the user (positive as well as negative). Therefore also data on the agent is personal data of the agents owner. An agent being just a tool legally seen also supports this.

Because of the last, an agent can also give consent for his owner with regards to the use of personal data. If the agent agrees where he should have not in the opinion of his master, it is his owner's fault for using a defective agent or having misconfigured it. The same is true if explicit consent is required, although the requirements for checking whether this is really the agent's intention will be higher.

Personalization Needs of E-Government Portals

A portal for E-Government should consist of the following areas:

- General information on the portal: No personalization is needed here, except perhaps help-pages (which can be handled exactly like the next category).
- Specific information on legal requirements and procedures: Special selection and hints to areas of possible interest are issues when personalizing these items. Helping decide which forms to fill in, where to go, what to bring with you, etc. ("wizards").

- Sending applications online (Filling in forms and sending them directly): Personalization can help in automatically filling in parts of forms and collecting additional certifications which are required and available online somewhere. Also merging multiple forms into a single one could be done (identifying prerequisites, consolidating it for presentation, separating it into different forms for different authorities).
- Online transaction (Online applications and also payment and/or receiving permits): Personalization is of less importance in this area, aspects of security and transactional processing require more attention. However, details on payment or secure identification for non-repudiable service (see recognizing users) can be stored in the same way as data for personalization: both must be protected. Agents are most useful in this area, as they can regularly check for new information or stages of proceedings and also automate any necessary payments. Identification and/or authorization can be securely proved by attribute certificates in this case ([Sonntag 2001]).

Requirements

Several key issues must be taken into account when discussing or implementing personalization for E-Government portals:

- Users have an extremely broad range of knowledge on computers, ranging from absolute novices to experienced experts. Also, classifying them according to this can be very difficult.
- Personalization must be very reliable: There might be liabilities if wrong advice or hints are given, especially as it is an "official" site. Even if not, bad word of mouth is also an issue if problems occur.
- Users will be extremely unwilling to provide information not already available to public administration. Even then partitioning the data and restricting its use to certain authorities (e. g. internal revenue service only) is important to them.
- Privacy contains a principle of minimalism: Only this data may be collected and used, which is absolutely necessary for the purpose on hand.
- Data must be kept up to date according to its purpose. This might require continual verification or adjustment with the authoritative source (with the problems of consent to the transmission of data).
- The administration, which is probably the operator of such a portal, is strictly bound to the laws (legal permission needed for actions) and also must observe the basic civil rights (the latter is less of a problem, as privacy is a basic civil right which also applies between citizens; a notable exception).
- Access to information should also be possible through agents, which is of special utility for companies, which can then automate search for information or preliminary stages of applications, as well as getting alerted on changes of special interest to them.

Recommendations for Recognizing/Identifying Users

For recognizing users a combination approach seems best: Identifying users through cookies (after obtaining permission for setting them, possibly implicit through mentioning this) and providing additional identification through webforms (as well as possibilities for log-out for those sharing a computer with others). For security reasons, especially with a look towards transactions, a secure connection should be used for identification (and afterwards). For those wanting no personalization/identification or just information, anonymous access should also

be allowed. As a fallback, encoding the identity in URL's could be used, but this is optional as it requires some work (and especially computer-time on the server).

If a more tight integration is used, transactions are common or agents are employed, electronic signatures might be used. In this case identification should also be possible through certificates using the secure communications protocol in addition to any custom protocols. This may be used only with explicit consent of the user (either general or on each connection): Identification through a certificate is different enough from cookies (very secure, additional information might be contained in the certificate, identification as a certain person instead of recognizing him/her as the same person, …) so a different handling is required.

Ideas for Adapting Content

Search path shortening is not suitable for an E-Government portal, as trees will be rather shallow and changing the order or location would be a huge problem for inexperienced users. Collaborative filtering is only suited for a very small area: If discussion groups or FAQ's are included, these could be automatically classified through ratings according to the (otherwise found) interests of users viewing as well as rating them. Apart from personalization rating can be used as a feedback for the creators of the content.

In a certain way statistical profiles are important: If a classification was reliably established through other means, predetermined profiles can be used for enriching it. This might be done either though statistical information or legal requirements (e. g. company → corporation taxes). These additional traits should be openly accessible for the user so he can remove certain of them he does not want or which are not applicable in his case. For optimum service to citizens also explanations why they were selected should be offered (self-explaining system).

Observing the behavior of the user is an additional trait, which should be used for regular visitors: An example could be a personalized "Hotlist" containing those pages regularly used. Also, changes in the behavior can be used as hints that some data is no longer valid and must be updated, removed or at least marked as "suspect", in this way fulfilling the obligation to assure the correctness of the data used.

This leaves questionnaires as the main means of obtaining data for personalization. As many questions as possible should be optional. Improving personalization later on should be possible through answering additional questions. Combinations with pre-created profiles and rules (see statistical profiles above) can further enhance the information on the user.

Conclusions

The success of the Austrian E-Government portal help.gv.at [help.gv.at] has sparked large interest in similar and improving projects (e. g. [eGOV]). For going on in the same direction, personalization is an important issue. However, in contrast to just providing information there are also more problems and drawbacks, like legal issues of privacy and reluctance of users to provide personal information. Still, the advantages outweigh the problems and personalization must always at least be considered.

We discussed different methods of recognizing users, from which cookies and identification through webforms over a secure connection are the most appropriate combination for portals with regard to human customers and certificates for agents. Adapting the content is a bit more difficult here than in E-Commerce, as higher standards must be met for correctness (agents: clarity) of the classification. Therefore explicit information through questionnaires, enhanced by rules, regulations and statistical data should be used to focus the information available to a personal view.

Privacy issues must be addressed when personalization is used, most important through obtaining the permission of the user for gathering, using and transmitting his personal data. Because of legal requirements and as a positive example, absolute and strict adherence to the rules is a necessity. There is no difference here if agents are introduced as another means of accessing the information.

Providing special services to or though (see e. g. [Theilmann/Rothermel 1998]) agents or taking them into consideration can additionally improve quality of the service for citizens and especially for companies. Anyway, some questions arise in this connection which must be solved in the future: How can the interaction between an agent and the portal be done to require least modifications and effort? How to effectively pass areas of interest, whose description might be difficult (transporting also the meaning and the intentions, resulting in a transfer of knowledge)?

Personalization is a logical step onward in the process of bringing government closer to the citizen, enhancing and in some cases perhaps replacing the need for assistance through personnel, freeing up resources for more non-standard and complicated issues.

References

[Bonifati et al. 2001] Angela Bonifati, Stefano Ceri, Stefano Paraboschi: Pushing Reactive Services to XML Repositories using Active Rules. WWW 10, 633-641, 2001

[Brandl/Mayer-Schönberger 1999] Ernst Brandl, Viktor Mayer-Schönberger: CPU-IDs, Cookies und Internet-Datenschutz. ecolex 1999, 366

[CookieSpec] Netscape: Persistent Client State HTTP Cookies. http://home.netscape.com/newsref/std/cookie_spec.html (12.12.2001)

[Dierks/Allen 1999] Tim Dierks, Christopher Allen: The TLS Protocol. Version 1.0. RFC 2246. http://www.pasteur.fr/infosci/RFC/22xx/2246 (12.12.2001)

[DSG] Datenschutzgesetz 2000 – DSG 2000, BGBl. I Nr. 165/1999

[DS-RL] Richtlinie 95/46/EG des Europäischen Parlamentes und des Rates vom 14. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr. Abl. L 281; 23.11.1995

[eGOV] An Integrated Platform for Realising Online One-Stop Government. http://falcon.ifs.uni-linz.ac.at:8080/eGOV/ (12.12.2001)

[help.gv.at] Ihr Amtshelfer im Internet. http://www.help.gv.at/ (12.12.2001)

[Jahnel 2001] Dietmar Jahnel: Datenschutz im Internet. ecolex 2001, 84-89

[Jennings/Wooldridge 1998] Nicholas R. Jennings, Michael J. Wooldridge: Applications of Intelligent Agents. In: Nicholas R. Jennings, Michael J. Wolldridge (Eds.): Agent Technology. Foundations, Applications, and Markets. Berlin: Springer 1998, 3-28

[Kristoll/Montulli 1997] David M. Kristoll, Lou Montulli: HTTP State Management Mechanism. RFC 2109. http://www.pasteur.fr/cgi-bin/mfs/01/21xx/2109 (12.12.2001)

[Mühlbacher et al. 2001] Jörg R. Mühlbacher, Susanne Reisinger, Michael Sonntag - Intelligent Agents and XML - A method for accessing webportals in both B2C and B2B E-Commerce; in: R. Corchuelo, A. Ruiz, J. Mühlbacher, J. D. García-Consuegra: WOOBS´01; Workshop on Object-Oriented Business Solutions Proceedings book, Budapest 2001

[POND] The Agentsystem POND: http://www.fim.uni-linz.ac.at/research/agenten/index.htm (13.12.2001)

[Sonntag 2001] Michael Sonntag: Improving Communication to Citizens and within Public Administration by Attribute Certificates. In: Maria A. Wimmer (Ed.): Knowledge Management in e-Government. KMGov-2001. 2nd International Workshop jointly organised by IFIP WG 8.3 & 8.5, University of Linz and University of Siena. Linz: Universitätsverlag Rudolf Trauner 2001, 207-217

[Sonntag/Reisinger 2001] Michael Sonntag, Susanne Reisinger - Important Factors for E-Commerce; in: C.Hofer, G.Chroust (editors): IDIMT - 2001; 9th Interdisciplinary Information Management Talks, Zadov (Tschechien); Universitätsverlag Rudolf Trauner

[Theilmann/Rothermel 1998] Wolfgang Theilmann, Kurt Rothermel: Domain Experts for Information Retrieval in the World Wide Web. In: Matthias Klusch, Gerhard Weiß (Eds.): Cooperative Information Agents II.Berlin: Springer 1998, 216-227 (LNCS 1435)

[Wenger/Probst 1998] D. Wenger, A. R. Probst: Adding Value with Intelligent Agents in Financial Services. In: Nicholas R. Jennings, Michael J. Wolldridge (Eds.): Agent Technology. Foundations, Applications, and Markets. Berlin: Springer 1998, 303-325

[Wimmer/Krenner 2001] Maria Wimmer, Johanna Krenner: Next Generation von One-Stop Government Portalen: Das Projekt eGOV. In Bauknecht, Brauer, Mück (Eds.): Informatik 2001. Tagungsband der GI/OCG Jahrestagung. Band 1. Wien: OCG 2001, 277-284

[Wooldridge 1997]: Michael Wooldridge: Agent based software engineering. IEEE Proc Software Engineering 144(1) 1997, 26-37