

Academic Senate Committee on Computing and Communications

Comments on CPHS Draft Policy on Human Subjects' Data

Version 5, November 29, 2004

The following are preliminary comments on the Policy from COMP. We will continue to gather input, and provide future addenda as appropriate. The draft policy was distributed to the Committee members and to four other faculty in EECS and SIMS (with specific expertise on the technical issues represented here) for their review, criticism, and additions. This preliminary response represents inputs from six faculty in total. You are also encouraged to read the raw faculty [comments](#) directly.

Summary

We believe that such a policy on the protection of subject's data should focus on three primary objectives: reducing the probability of *human error*, using strong *technological barriers*, and minimizing the *impact* on research projects in both the near and longer term.

- In a situation where many researchers in disparate groups with varying levels of computer expertise are handling subject's data, human error is the primary security threat that you should be concerned with. Approaches that minimize the possibility of human error should focus on professional system management and administration, minimizing the number of places where data is stored and accessed (a server based solution), and avoiding the need for conscious actions on the part of users (automation). These go together—automation of security mechanisms and processes depends on expert installation and setup of software, and keeping the data in one place greatly reduces the opportunities for human error.
- The strongest tool in your security arsenal is strong encryption (by that we mean both adequate key sizes and proper management and distribution of secret passwords or passphrases). Subject's data which is conveyed or stored in encrypted form should not be of concern, even if inadvertently or deliberately divulged publicly. When encryption is properly used, nothing is accomplished by disconnecting a computer from the network.
- Solutions that minimize the impact on researchers and their projects rely on readily available commercial software (ideally available for a variety of platforms) or services (available over the network, and managed either commercially or by the campus central computing staff), or both. Although they

do require some effort to set up initially, they should not require day-by-day conscious actions on the part of users (for both security and operational reasons) or special expertise during the course of usage (as opposed to installation and configuration). They should allow researchers to use the tools they are familiar with (Excel spreadsheets, statistical software, etc.).

The current policy statement relies on two primary tools: de-identification of the data and disconnection from the network. We believe that de-identification is an excellent idea, especially because it will reduce the opportunities for human error (by segregating sensitive data in separate files that are presumably infrequently accessed and used). Disconnection from the network, on the other hand, is not a desirable tool from any of these three perspectives. It requires day-by-day conscious action on the part of users, as well as a level of expertise, and leaves many opportunities for human error that would expose data. It does not employ encryption, and thus the subject's data remains vulnerable to theft or accidental disclosure. Disconnection from the network will cause all sorts of operational challenges, both in the administration of the systems and in the conduct of the research itself.

Fortunately commercial software and services are available (and hence can be deployed quickly and relatively inexpensively) that can provide a high level of data security. There are three main categories of approaches that we have identified: Distributed storage of sensitive data in encrypted form, storage of subject's data on a secure central server, and management of subject's data in a central database management system (DBMS). We believe the DBMS is the superior long-term solution, but recommend local and/or server encryption as a transition strategy—it is quite secure (considerably more so than disconnection from the network), transparent to applications running on the same computer where the data is stored, and allows for secure backups or secure storage of sensitive data on removable media.

Characteristics of a "good" security architecture

Before examining the draft policy document it is helpful to review in a little more detail the criteria by which a security architecture should be evaluated. Many of these points are elaborated on below.

- Security technologies are available that provide very high levels of confidence. However, they have to be deployed and used properly to be effective. Thus human error is *the* major problem that technology cannot fully overcome, and human malfeasance is the second most serious problem. The single most important lever is to minimize the opportunities for human error, and to insure that (in the absence of human error) malfeasance can only occur in a context covered by Federal and state criminal statutes.
- As few people as possible should have need or authorization to access the data. Those people should be strongly authenticated before allowing access.

- Any policies that require a human being take specific actions every time they deal with data is inherently susceptible to human error. The most effective security technology is automatic and transparent, working silently behind the scenes.
- Security breaches should be visible, so that corrective action can be taken as soon as possible.
- Since adhering to technical requirements requires some technical expertise in the course in installation and configuration of security measures (and in the case of less desirable solutions during usage as well), all machines on which sensitive data is stored however briefly should be professionally managed and administered.
- The fewer computers store or process sensitive data the better, because this reduces the opportunities for human error. Countering this is the observation that where breaches do occur they are potentially more extensive. Hence the next bullet.
- Sensitive data would be strongly encrypted while it is stored and while it is in transit from one computer to another. Ideally it should only be decrypted dynamically while it is being processed, and only in volatile memory (so that a decrypted version cannot be left behind inadvertently).

Generally these observations favor a centralized storage architecture, where only one or at most a few secure servers are the focus of secure data storage and access. Conversely, an architecture where data is stored in a distributed fashion across many computers, many of them personally owned and not professionally administered, is undesirable, and especially where some of those computers are not professionally managed. The alternative architectures we describe below are chosen with these observations in mind. However, we make no representation that these are the only feasible solutions, only that they are feasible and desirable ones. First, we make a number of comments on the current draft of the policy.

General comments on the policy (independent of the specific technical approach)

1. In our view, one goal of the policy should be to maintain the continuity in the research activities while transitioning to a new more secure environment. We believe that the draft policy would not achieve this goal. Without a transition period, the ‘cost’ of imposing this policy will be much higher (delay, idle researchers and staff, etc.) and the benefit-to-cost ratio significantly and adversely impacted. On the other hand, we believe that the security of data during a transition period can be dramatically improved without imposing highly burdensome requirements. We describe one feasible approach in alternatives one and two below.
2. Why wouldn’t it be acceptable to couple the timing of the *final policy* implementation to the availability of feasible software and hardware solutions, plus a reasonable (but short) period to learn the new rules and mechanisms and move the data into its new environment? Since security breaches have been

isolated and infrequent, the probability of a security breach during this transition period would be very small, and could be made much smaller with some simple measures. We therefore suggest separate *transition* and *final policies*. Several specific suggested alternatives for a transition policy:

- A requirement that encryption software be installed on each computer storing sensitive data. (See a description of this option as alternative one below.)
 - A requirement that sensitive data be stored in a commercial secure storage facility for all periods that it is not being actually used would be simple and cheap to implement. (See a description of this option as alternative two below.) This would have the weakness that sensitive data would be stored on a campus computer during periods when it is being actively used.
 - Researchers could be instructed to take immediate and substantial steps to secure sensitive data by means of their choosing (accompanied by clearly stated goals and a list of suggested approaches, such as our alternatives one and two and three) with a requirement to report what steps they have actually taken (or why they feel their security is already strong) to CPHS on a form provided for that purpose.
 - Separate transition policies for cases where sensitive data is stored on a professionally managed computer and it is not. In the former case, researchers would be required to either move the data to a professionally managed computer, or to store it on a removable medium in a physically secure location. In the later case, they would be instructed to take immediate steps to secure the data by encryption as suggested previously.
3. If you do intend to impose a policy (transition or final) without delay, we at least strongly recommend a prior period (perhaps short) for public comment. We strongly suspect that many in the research community will find ways to fault the policy that cannot be anticipated, you will learn a great deal about the impact of the policy, and they will be more understanding and cooperative if afforded an opportunity to air any concerns.
 4. The policy should begin with a statement of specific goals, such as “data stripped of identifiers can be publicly divulged” and “personal identifiers must not be revealed to any but authorized researchers for purposes identified with and related to the research objective”.
 5. We recommend that the policy be accompanied by a concise white paper on the general considerations in security and what measures should be taken. This could be a vehicle for educating the community as to their responsibilities and the tools

at their disposal. We also suggest that all administrators and users of systems storing sensitive data be required to attend an orientation course supplied by CPHS on protecting data—goals, ways to reduce human error, and required technical measures.

6. There needs to be a clear definition of what data is covered. To take an extreme example, consider an electronic version of a telephone book. These data certainly contain "personal identifiers". Presumably the telephone book need not be kept under lock and key. Or, to take a more germane example, consider data obtained from a public agency including names, addresses, and telephone numbers of holders of licenses to purchase wholesale agricultural commodities, along with some limited data on the characteristics. These are public data, in the sense that the state of California will provide them to any interested party, though these data can't be found online. Should such data be covered by the CPHS policy? Should any data obtainable via a FOIA request be covered (e.g., salaries of UC faculty)?
7. "qualifications of the individuals..." This is extremely important, and in our view greater prominence should be given across the board to the need for professional administration and management of any systems storing personal identifiers, the qualifications and training of those individuals, and the training of all individual users of the system. Such systems will always have system managers and administrators (and in some cases this may be the user), and those individuals may have unconditional access to the data and they therefore have to be trusted (this is especially true if there are recovery policies to prevent loss of data). You should note that security breaches are very rarely the fault of the technology involved, and almost always the result of human error. Your *primary* leverage point for limiting breaches is therefore to reduce the incidence of human error or malfeasance, and there are two ways to do this: reduce the opportunities for error or malfeasance (by limiting the number of people involved and limiting their level of activity) and reduce the probability of error (by making the security measures automatic) or malfeasance at each opportunity (by measures to insure that system administrators are trustworthy, reliable, and competent).
8. For reasons stated in the last comment, we strongly suggest a requirement that any system storing sensitive data (even temporarily) must be professionally managed and administered, with a vetting and approval process for his or her qualifications. If this is achieved, and if the goals are clearly communicated to this person as suggested above, you are almost home free. An obstacle here is that some departments cannot afford (or at least have put a low priority on) technical assistance to faculty, staff, and students. We believe this issue must be addressed, at least where sensitive data is involved. A central recharge system administration service operating on recharge would be an efficient way to address this need.
9. The policy should explicitly state that in all instances where personal identity information (or a portion thereof) is not necessary to meet the research objectives

then such identity information should be both stripped and *permanently discarded*. In our view after this is accomplished all security requirements can be waived. (This will also provide compelling incentives to strip and discard.)

10. We do agree that minimum technical standards are necessary. Complying with general campus base level security requirements should certainly be a part of that. But by itself this is inadequate, because sensitive data needs to be stored in encrypted form and this is not part of the campus minimum security policy.
11. When it comes to specific security architectures (such as your 'non-networked computer') we believe that it is overly restrictive to impose a *single* solution. Rather, we suggest that there be a list of acceptable security architectures, together with a process for vetting and approval of new architectures added to the list (using campus or commercial security expertise). In fact, all architectures on the approved list should be *proposed* by, not just approved by, security experts. The list may initially be very short (one or two), but could grow with time and experience. We suggest that you write the policy with this structure from the start, even should the initial list be short.
12. We believe that *any* configuration that offers data security that meets or exceeds certain standards (in an expert's opinion after external review) *and* which is professionally administered should be acceptable under the policy. This could be implemented by setting up a vetting and review process by which data security plans not on the aforementioned pre-approved list can be vetted and approved. We would expect that most projects would choose a pre-approved approach, so this would not impose a high administrative burden.
13. Emailing of data with personal identifiers is a *very* bad idea unless data itself is encrypted using strong encryption. Email is so insecure that this is tantamount to posting the data on a public web server. There are viable and easily available alternatives:
 - First, it can be mandated that only encrypted files can be mailed. This does require that the recipient has the appropriate decryption software and that there be a secure way to convey the password or passphrase (for example, through a telephone call would probably be adequately secure). It is also important that a *different* password or passphrase be used for email transfer than for local storage. All this offers considerable opportunity for human error, but our alternative one below offers a viable and relatively simple solution.
 - There are good alternatives to email for transferring data securely from one computer to another securely. For example, [SSH](#), a secure file transfer program with a campus site license can be used, although again it requires that the destination computer have an SSH-compatible server.

14. It should be noted in both the policy formulation and in the policy itself that files that are deleted are typically not actually erased from the disk unless a "disk wipe" program is used. Thus, it is possible for this data to be recovered later (much later), for example if the machine is stolen. Fortunately, we believe that campus excess and salvage routinely performs a disk wipe on discarded machines for exactly this reason (this should be confirmed), but what about personally owned machines? This is another example of where there are multiple opportunities for human error, and where professional system administration is important. The policy should specify that machines storing sensitive data in unencrypted form, even briefly, must be wiped regularly. (Because this is another opportunity for human error, this is another argument for a solution that does not require unencrypted data ever be stored on a local machine, such as our alternatives one and three below.)
15. We do not understand why there is any justification for imposing security requirements on 'de-identified data'. In fact, would not the scholarly process normally encourage researchers to make this data available to others for verification of results or for follow-on research? If personal identity information cannot be inferred from the data, why cannot the data be publicly divulged? On the other hand, we can imagine rare circumstances where personal information could be inferred implicitly rather than explicitly, such as a study that had only one subject. Thus, we suggest that data be exempted from this policy entirely if (a) no personal identity information has been collected or such identity information has been stripped (that is de-identified) *and* (b) there is no feasible way a determined person could extract identity information from the data by other implicit means.
16. We anticipate that one impact of this policy will be to drive researchers toward more paper based processes. But they should recognize that where that paper is printed from a computer, the data necessarily resides in a computer (say in an Excel or Word file) and identical security issues arise. Thus, they can avoid these policies only by *handwritten* paper records. Generally the policy should deal more explicitly and completely with paper records.
17. We suggest that CPHS itself should at all times have at least one member who is an expert in IT and familiar with data security standards and technologies.
18. Computer security is a constantly changing area. Thus, this policy should be reviewed and updated on a regular basis. Each update of the policy (and the current draft) should be vetted with computer security experts (in IS&T, on the faculty, or outside consultants).
19. To a person of computer persuasion, the term 'hacker' has a positive connotation as someone interested in technical challenges without evil intent. A better term for

someone who breaks into systems for nefarious purposes is 'cracker' (defining the term the first time it is used).

Specific comments on the policy draft

The policy proposes a single security architecture for storing sensitive data—the non-networked computer. This section gives comments on this specific technical approach.

20. The proposed policy is not very secure (because it does not employ encryption), is relatively vulnerable to human error (because it requires ongoing and explicit actions on the part of technically unsophisticated users), and in our opinion would cause many operational problems and expenses (because it relies on non-networked computers).
21. We are concerned about the strength of security of using non-networked computers (even if password protected) or non-removable storage that is physically locked up. This approach perpetuates (and likely even proliferates) machines, many of them not professionally administrated, and thus creates more rather than less opportunity for human error. Burglarizing of campus labs and offices is fairly common. Any thief who gains physical access to a computer, even one that is password protected, can access the data stored on that computer (for example by booting it up from a removable disk). Burglaries of computers without the express intent to steal data is equally bad, because often these computers are sold to third world countries or in flea markets, and identity theft thieves specifically seek out these machines. In summary, from a strength-of-security standpoint, we would discourage this approach from being included in a list of acceptable security architectures.
22. We believe that imposing this policy immediately and universally as written will carry a high cost, because as a practical matter most human subject research projects that must retain sensitive data would have to shut down for a matter of weeks (at least) while they implement the new mandated measures and acquire the hardware and software necessary to do so. In most cases they will not be able to use existing equipment for storing personal identity information (as such equipment is typically used for multiple purposes, many requiring network access) and will have to acquire a new computer for this purpose. Requiring many research groups to acquire a new computer and professionally administer it is also expensive overall; we suggest alternative solutions with stronger security below that would be less invasive.
23. Requiring non-networked computers for sensitive data is also burdensome on an ongoing basis. It is fair to say that computers without network connection are rare today, and that most existing computers on the campus have multiple uses and require network connections for many of these functions. Even a computer which is used exclusively on a given project needs to have software patches and upgrades and be backed up, and this is commonly done through a network

connection. Thus, we believe that many research groups would have to acquire an additional non-networked computer for this purpose, and face almost insurmountable problems in administering such a configuration (in terms of patches, upgrades, remote administration, and backup, among other issues).

24. Not having remote or network access to this data would in many cases be burdensome to the researchers themselves, because many normal procedures for managing project data involve the network. As an example, this policy would completely eliminate the use of the internet for data collection (such as online questionnaires) in a study that falls under the jurisdiction of the CPHS. Because the computer that collects the data is by definition connected to the internet, and the original data must be stored, at least temporarily, on that computer it violates the letter of the policy. Reverting to paper-only questionnaires would not even be acceptable if, for example, scanning technology was used and the scanner was network connected. As another example, many projects use centrally supported statistical analysis packages, and this would be prohibited. As another example, perceptual experiments in psychology involve repeated measurements (for example, of vision or audition) from three to six observers over a long series of days (often months) and in which the results of one day must be used to set up the next. These rules would greatly encumber such research. As another example, much social science research involves collecting spoken interviews and sending them by network to contract transcription services. This would be prohibited.
25. There are encryption mechanisms that can protect the data stored on a computer or removable disk from attack (even on a compromised physical location), but the policy does not mention or require them. This is a major omission. In fact, if strong encryption is used properly there is no need to disconnect a computer from the network nor to worry about physical security of either a computer or removable medium.
26. Secure remote access to personal identifier data is technically feasible using available commercial software. As evidence of this, note that brokerage houses and e-commerce sites (like schwab.com and amazon.com) offer networked access to personal information. Overall, the number of such sites must be at least in the thousands, and they have existed for a decade or so. The campus has considerable experience in running similar applications, for example e-grades, e-commerce, selling athletic tickets, and many others. The only breaches of private information that we are aware of on these sites were the result of human malfeasance—'insiders' stealing data and selling it. But any security architecture is vulnerable to insider malfeasance, which is punishable under Federal criminal statutes. It is instructive to examine the reasons that these server solutions offer strong security:
 - Their operating system is more secure, having been designed with security in mind from the start (because they are multi-user servers)
 - Invasive applications with security holes (email, Web browsers, etc) are kept out

- Additional security measures like firewalls, intrusion detection software, etc are used
- They are kept in highly physically secure locations
- They are professionally managed and administered, with greater staffing resources

If you do wish to impose this security architecture in spite of these negative comments, the following are comments on the policy as it is written:

27. One alternative is to store this data on recordable media, but if that media cannot be read and the data on it processed on a networked computer, we are back to a burdensome requirement for a non-networked computer in each research group. Where data is stored on a removable medium, the policy is not clear as to the security requirements of the computer where this medium can be read and used. It might be inferred—reading between the lines—that this computer cannot be networked. For the aforementioned reasons, this would be extremely burdensome, and provide little in the way of advantage over storing the data in non-removable storage.
28. Your permission for de-identified data to be emailed is a far looser requirement than the requirement that it be stored on a password protected computer. Emailing such data is not that much different from putting it on a publicly available web server, as email is *highly* insecure. We thus reluctantly recommend that either (a) emailing of even de-identified data not be allowed, or alternatively (b) that no specific security requirements be imposed on a computer containing de-identified data. Note that there are viable options to emailing of data that *are* secure, including a removable medium or file transfer using a secure application like “SSH Secure Shell” (note that is application is already available for free on the campus software distribution web site). It is also secure to email encrypted files. Thus, the best policy option of all may be (a) emailing is prohibited but (b) transfer by secure file transfer software (and list acceptable options) is permitted. But if there is literally no concern about public disclosure of this data (which is unfortunately consistent with emailing it), then why impose any security standards at all? As noted above, we advocate not placing *any* security restrictions on that type of data.
29. “Non-electronic data on paper, audiotape or videotape” Note that audiotape and videotape *are* electronic media, and also that audio and video can reside in non-removable computer storage and well as on removable magnetic tape.
30. “sometimes it is not practical to remove identifiers”. It should almost always be possible to ‘join’ de-identified data with identity information dynamically in the course of processing the data (using the key code as the linkage). Doing a separation of data from identity is a useful step for maintaining security, but this ‘join’ operation becomes virtually impossible if the identity information resides on a non-networked computer. This is one example of where aspects of the policy

work at cross purposes, or in other words where security would be *enhanced* by in fact having the identity information available for secure access across the network. It would be helpful to give examples of where it is not possible to separate identity information. The only example *we* can think of (assuming the aforementioned 'join' is feasible) is audio or video, where the subject may be asked to state their name or other identity information. Also, this paragraph addresses a new project application to CPHS, but what policies apply to existing projects?

31. A 'non-networked computer' is defined as one that is 'not connected to the Internet'. Even if this non-networked requirement is included in the policy (and as stated above we believe this is neither necessary nor strongly secure), the definition of 'non-networked' needs to be much more carefully defined. For example, we believe that security would not be compromised (relative to the non-networked case) if a set of computers co-located in a physically secure location were connected by a LAN, if that LAN was in turn isolated from the public Internet. In fact, such a configuration is an integral and unassailable part of any 'cluster computing' configuration which may be used in some projects. Calling such a configuration non-networked will lead to some confusion in policy implementation. If you follow through with this requirement, we therefore strongly suggest that you give a list of acceptable networking architectures (a list that can grow over time) together with a vetting and approval procedure for other configurations.
32. "absolutely necessary to hold data which has not been de-identified on a networked computer". We can easily imagine that reasonable people will disagree on what constitutes 'absolutely necessary'. Thus we believe that both a definition of the term and a vetting and approval process (or at least a reporting process) are necessary. <<We plan to look into the stated campus security policy to judge whether it is in fact adequate. For example, we would not judge it to be adequate if it does not include encryption of data on access and some strong authentication.>>
33. "CPHS requirements are stronger..." This whole paragraph requires considerable refinement. For example, it is not clear why a network connection might be 'absolutely necessary', but it strikes us that the only way this high bar could be exceeded is if it is necessary to access or process the identity data over the network. (If it is only so the computer can be used to send and receive email, then this seriously undercuts the whole policy, because *everybody* can and will invoke this rather weak justification.) But then you say that the configuration must prevent any such remote data access, legitimate or otherwise, but this seems to contradict the premise. In addition, it is unlikely that a firewall alone could ever, by itself, meet this requirement, but the discussion is limited to firewall configuration (encryption is far more important). In our judgment that the requirement that the computer be networked disconnected most of the time is not

going to do much good, and will in addition be burdensome, because either the computer is secure or it isn't—there isn't a lot of gray area.

34. “exceptional circumstances...” We do not believe that it should *ever* be necessary to email data with identifiers and without encryption, because there exist viable and widely available alternatives (like SSH) or physical transport of a removable storage medium or encryption of files before being transferred, and as noted above email is very insecure. We therefore strongly advise against allowing this exception. On the other hand, we believe that remote access to this data *is* a compelling need in many cases, and can be accomplished securely (as in our alternative three below), so we suggest that this not require an exception but be permitted under a revised policy with appropriate precautions as we have stated above.
35. We do believe that allowing exceptions (with approval) is very desirable, because there will inevitably arise circumstances that cannot be anticipated in any policy, no matter how carefully crafted. P. 1, paragraph 5: CPHS talks about allowing exceptions to the rule of removing all identifiers. However, in paragraph 2 of the same page we're told that the policy takes effect immediately and applies retroactively. How is someone who has affected data and already has CPHS approval supposed to obtain an exception?
36. “Requirements...” This paragraph seems to presume that data is stored in a ‘file’. In fact, in many cases the data will be stored in a commercial database, and that database will in many cases management multiple files. (In fact, in our view this is far preferable from a security perspective, since databases include many access control features.) We can imagine circumstances where the data might be stored in some other (possibly homegrown) application. So this whole set of requirements should be revamped to anticipate a broader range of circumstances.
37. Simply turning a computer off is usually preferable to disconnecting it from the network (and saves energy in addition), unless the computer can be used for some other purpose while disconnected. Thus we suggest the policy treat turning off a computer as equivalent to disconnecting it from the network.

We will now review several security architectures that would offer an alternative to that proposed in the policy.

Alternative one: local encryption

Numerous solutions exist for the encryption of files stored on a computer or on a removable medium. Once a file or folder is marked as encrypted, then it is stored on disk in encrypted form even as it is processed in memory in unencrypted (clear) form. That way, if a computer is broken into over the network, or stolen and physical access to the computer is obtained, the perpetrator will be unable to access the content of the encrypted files (unless he or she is able to learn the password or passphrase). Similarly, this

approach applies to removable media—the files on those media can be similarly encrypted and would thus prove useless if the medium is stolen.

This capability is built into some more recent operating systems (for example the Encrypting File System (EFS) in Microsoft Windows XP Professional), and is available as both public domain and inexpensive commercial software. We do not recommend depending on EFS because it is not available on older versions of Windows and because it has a number of shortcomings. Many applications (like Word or Excel) also have password protection features, but we do not recommend using them because they are easy to crack with free software available on the network (they are designed to prevent inadvertent eyes or manipulation of a file, rather than deter a serious cracker).

As an illustration of what *is* possible, we investigated the products from [Cypherix](#) for the Windows operating system. They have a family of products called Cryptainer that works with all 32-bit versions of Windows (95 through XP). This utility creates an arbitrary number of 'secure vaults' on a computer. Each secure vault appears to Windows as a mounted disk, and up to four of these secure vaults can be mounted at any time. For example, [Cryptainer LE](#) is free, and limits each secure vault to 25 MB of data, and other versions available at modest cost allow larger vaults (up to 50 GB). Secure vaults can also be created on removable media (like a Zip drive or USB drive). Each vault is protected by a passphrase (e.g. a nonsense sentence that is easy to remember) of up to 100 characters, and vaults are virtually impossible to crack using available techniques and computers (they use encryption algorithms sanctioned by the US government for its internal use, and encryption keys with a length beyond the capability of computers for the foreseeable future). In order to mount any secure vault as a disk, its associated passphrase must be produced. Once it is mounted, any program running on the same computer within the same user account can access files within the vault transparently, as if it were on a regular (unencrypted) mounted disk. As the program is running, each time it requests data from the vault, that data is unencrypted in volatile memory as it is read, and any time the program stores data on the disk it is encrypted as it is stored. Thus, the data is never stored on disk (or removable medium) in unencrypted form during normal operation.

The Cypherix products, when used properly, only store files on disk (permanent or removable) that are encrypted and are unbreakable by anyone not knowing the passphrase. There are limitations to this security, however:

- A program running on the same computer and under the same account as the user who mounted the secure vault can pass an unencrypted version of the data on to another computer or user. This is because the data in any vault that is mounted as a disk (which requires producing a passphrase) is freely available to any program running in that account. This may create some opportunities for a cracker—we would like to see further investigation of this.
- A user can inadvertently copy a file from a vault mounted as a disk into some unencrypted area of the local file system, thereby leaving an unencrypted version of a file around for later theft.
- A user who leaves herself or himself logged in with secure vaults mounted leaves the content of the secure vault available to anybody with physical access to that

machine. Thus, it is necessary to manually de-mount the secure vaults or logoff the machine to be assured that the sensitive data is secure against physical access to the machine or removable medium.

- As with any encryption scheme, it is vulnerable to the theft or inadvertent disclosure of the passphrase. For example, there exist Trojan worm programs that transmit all keyboard keystrokes to a cracker, and such a cracker could learn passwords or passphrases. To counter this threat, the computer should be professionally managed and security software (such as Symantec Client Security, for which the campus has a site license) should be installed.

Cryptainer also allows encrypted forms of files to be emailed securely. For this purpose, it creates a self-extracting file that contains not only the data but in addition a Windows program that can decrypt the data. The recipient does need to know the passphrase, but does not need to have purchased or installed the Cryptainer software.

Plusses in this approach:

- Very inexpensive
- Immediately available, easy to set up
- Easy to use—secure vaults appear to the user just like any mounted disk, but are actually contained in encrypted files that can be backed up to removable media securely.

Weaknesses and disadvantages in this approach:

- Requires conscious action on the part of the user or system administrator (such as creating a secure vault and then using only that vault for storing sensitive data files) This is an opportunity for human error.
- The variety of platforms used on campus makes it difficult to adopt a uniform solution (although we believe a solution will be available on virtually every platform).
- The most serious disadvantage is that it is possible to lose data because a user forgets the passphrase, or leaves employment, and so forth. More sophisticated solutions, but ones that are more complex to set up, have a recovery mechanism that allow one or more designated 'recovery agents' to decrypt files should the original user be unavailable.
- Data can also be lost because it is not backed up, a disk crash or theft, etc.
- Unlike our alternative three below, this option does not by itself offer an option for secure remote access to sensitive data.

This approach has significant advantages over that proposed in the draft policy. It is considerably more secure (not vulnerable to theft or physical access to the machine), considerably less vulnerable to human error (automatic after an initial setup of a secure vault), and less invasive (because the main motivation for disconnection from the network is removed and this should be unnecessary). This and the next alternative are attractive as a transition solution because of their low cost and immediacy.

Alternative two: networked secure storage

There are many commercial services that offer secure data storage and access over the Internet at modest cost; see for example [Xdrive](#). The way these services work is that they provide a client program that encrypts the data before it is transferred to the server for storage and after it is retrieved from the server for processing. Thus, the data is never available in clear form outside the client computer (passing through the network or as stored remotely). A user would normally store sensitive data files on the remote server. When it was necessary to access and use the files, the user would copy the files back to their own computer, where they would be stored temporarily. This transfer is easy to accomplish, appearing to the user like copying files within the computer itself.

It would also be possible for the campus to deploy its own secure server, rather than utilize a commercial service. This has the advantage that the standard campus authentication mechanism CalNet would be used as the basis for accessing and decrypting stored files.

It is likely that most secure storage servers require the user to create a local unencrypted copy of a file before it is used as we have described. This would be less secure than alternative one (which normally does not create an unencrypted copy of a file). There may be services that allow the remote files to be mounted for direct use by local applications, in which case it would work similarly to alternative one but would be lower performing. It has the advantage that the remote server will always be frequently and professionally backed up. Otherwise, this alternative has many of the plusses and minuses of alternative one.

Alternative three: A campus secure data warehouse

The following solution is more sophisticated and, we believe, offers compelling advantages as a long-term solution. It would store sensitive data in a secure server based on a database management system. At the expense of greater effort in initial setup (which is why it is less suitable as a transition solution), it has the compelling advantage that data would never be stored in unencrypted form (on either client or server machines) and all the security features would be automatic (greatly reducing the opportunities for human error).

This solution assumes the campus supplies a centralized secure data warehouse based on commercial database technology where sensitive data can be stored and accessed, where access utilizes encrypted transmission (using available applications like Web browsers equipped with SSL or SSH, see below), CalNet authentication of the user, and auditing and logging of all accesses. This warehouse would actually be considerably more secure than the approach mandated in the policy, because (a) CalNet is a strong and secure authentication mechanism and (b) the campus data center is physically *much* more secure than faculty offices and research labs (for reasons stated in 4) and (c) this centralized approach greatly reduces the opportunities for human error or malfeasance, as would its automation of security functions like encryption.

All major database management systems (Oracle, for example) have available strong encryption for data, and secure connection technology. This does have implications for normal DBMS management practice, such as additional security restrictions on handling backup and recovery datasets, etc. Thus, expert database management and administration is necessary, but this would be focused on a single machine, and the perpetrator of any malfeasance could more readily be identified. (This is why armies have forts rather than distribute their soldiers across the countryside.)

Unlike a file storage solution, the database management system presumes that the data has a structure (records and fields). On the plus side, this structure can be exploited for a wealth of standard-feature server-side processing functions (such as the 'join's mentioned earlier), and on the negative side many campus groups would have to translate their data from some other format. For this purpose, the campus could provide a recharge consulting service associated with the data warehouse. In the long run, research groups would benefit in many substantial ways from having their data accessible in a database, which has a wealth of features and capabilities.

At the expense of some initial setup, in most cases using commercially available software, the user's client machine can access and modify data in unencrypted form without storing in on disk. For example, Microsoft Excel supports an Open Database Connectivity Driver (ODCD) for all commercial DBMS servers. The data in the remote server can be processed locally by Excel in a manner that appears to the user exactly as if the data were stored on the local disk in unencrypted form (but it isn't). Thus, the possibility of both data theft and human error are very low. Of course, it would be possible for a user to capture unencrypted data and store it on a local disk (for example in a cut and paste operation), but this would require conscious and deliberate action on his or her part.

This is similar to a number of applications already running on campus, so it would be relatively simple to setup. The costs of the data warehouse would be low, and could be recovered by recharge. Bundled staff services could assist researchers in capturing the data to compliant formats and more generally adhering to these policies.