

# REPORT ON LONG-TERM ELECTRONIC ARCHIVING (LTEA)

## CONTENTS

1. Background & Purpose	3
2. Method	4
3. Results	5
3.1 General features of LTEA	5
3.2 CERN-specific implementation aspects	5
3.3 The present archiving situation at CERN	6
3.3.1 Description by storage environment	6
3.3.2 Description by document type, structure, complexity	7
3.3.3 People responsible for documents	7
4. Recommendations: Towards an affordable LTEA	8
4.1 Concrete actions	8
4.1.1 Actions to be taken in the area of electronic mail	8
4.1.2 Actions to be taken in the area of normal web pages	9
4.1.3 Actions to be taken in the area of CISs	10
4.1.4 Actions to be taken in the area of databases	10
4.1.5 Actions to be taken for paper-based documents	10
4.1.6 Actions to be taken for documents in "other areas"	10
4.2 Preparatory measures: A CERN-wide document-handling policy	10
4.3 Accompanying measures: foster meta-data	12
4.4 Actions and decisions to be postponed	14
4.4.1 General reasons	14
4.4.2 CERN-specific reasons	14
4.5 Authors motivation	15
5. Summary of Recommendations	17
6. Conclusion	18
Annexes:	
1. Composition of the WG	19
2. CERN Certified Information Systems (CIS)	21
3. E-mail long-term archiving	25



## **1. Background & Purpose**

In the context of world-wide growing awareness of the volatility of electronically stored information, the DG initiated in June 2000 a working group (see the group's composition as Annex 1) with the mandate to produce a set of implementable solutions in the field of Long-Term Electronic Archiving (LTEA) at CERN. We summarize the main purpose of LTEA as follows.

LTEA is needed in order to:

- a. provide the upper management with data to be considered in decision taking,
- b. provide proof that CERN is fulfilling its role and obligations,
- c. provide for posterity a record of the history and activities of CERN.

The sequence of these items reflects the time scale in which the data are likely to be consulted (first item = shortest term).

## 2. Method

The working group used the following procedure:

*Meetings:* The working group had 9 meetings, in which all general issues and some CERN-specific questions were discussed.

*Study of work done previously:* The results of the work of the previous working group, especially those concerning the situation at CERN, were examined and taken into account.

*Study of "literature" (mainly web):* most of it was provided by Anita Hollier and Corrado Pettenati.

*Study of EDMS and CDS:* interviews were held with representatives of EDMS and CDS, two existing document-handling systems at CERN, which almost fulfil the conditions of the so-called Certified Information Systems (CISs).

*Discussions with the CERN Legal Service:* these mainly concerned the problems of keeping legally sensitive documents in electronic form only.

*Discussions with companies:* A subgroup was formed to study the possibility of outsourcing the archiving of CERN's web pages as a whole. Detailed discussions were held with several companies regarding the feasibility and the cost of such archiving.

*Presentations to the Archive Committee:* The project of web archiving was presented at the 37th Archive Committee meeting, and the outline of the complete report at the 38th meeting. Valuable feedback was obtained at both meetings and the suggestions were taken into account.

### 3. Results

The methods described above allowed us to divide the complex task of designing LTEA solutions into four main subtasks:

- describe the general features of LTEA,
- describe CERN-specific implementation aspects,
- evaluate the present situation at CERN,
- elaborate recommendations for a general long-term archiving policy.

#### 3.1 General features of LTEA

The list below presents the general features that must be considered as independent of any particular case of LTEA:

- context
- meta-data
- content
- internal complexity of the document
- original format
- archiving format
- rules determining which documents are retained and for how long.

#### 3.2 CERN-specific implementation aspects

In order to have a realistic implementation plan, we have identified the key factors that determine the degree to which the LTEA implementation can and should be done. We have also identified the criteria for prioritization.

The key factors for a successful implementation are:

- resources (for buying or building and maintaining the system, entering of documentation),
- motivation of people directly involved,
- attitude of middle management (level of division, group),
- expected usage of documents (e.g. historian or lawyer),
- state of document *before* archiving,
- relevance of the document for operational use at the moment of saving and shortly after,
- selective or bulk storage,
- access rights, confidentiality.

#### *Aspects of priority*

The following criteria can serve as guidelines for the establishment of priorities and the sequence of actions:

- the area in which potentially valuable information is lost,
- the actions that would, if postponed, increase the cost of the final system,
- the actions that could lower the cost of the final system without harmful side effects

- the ease and cost at which a feature can be implemented,
- the preconditions for in-house or outsourcing LTEA.

### 3.3 The present archiving situation at CERN

Three very relevant and very worrying statements can be made from the outset:

- at this moment we are constantly losing potentially valuable information;
- there is no CERN-wide document-handling policy;
- unless such a policy is formulated and implemented, LTEA cannot be achieved or it could only be achieved at a prohibitive cost.

In order to assess the situation in more detail, different "storage environments" used at CERN have been analysed.

#### 3.3.1 Description by storage environment

##### *Electronic mail*

The use of e-mail as a very easy to use and widely accessible means of storing information must not be underestimated. The senders of e-mail do not necessarily have the intention to store information, but it turns out that almost every mail user, at some point, will search through his/her folders to retrieve information that would otherwise have been lost. This method of storing information is shaky, as it depends entirely on the behaviour of the individuals involved in the exchange of information. Once the individual has left CERN or, for some reason, "cleaned" up the folders, the information is lost. What is done at CERN at the moment in the field of archiving of e-mail can be summarized as follows:

- all mail sent to so-called listbox lists is archived,
- nothing of this kind is done with mail sent between individuals,
- there is no archiving scheme for mail folders of individuals.

##### *Normal web pages*

Storing information on the web is almost as easy as in e-mail. It is always done with the intention to "publish" information and is in many cases the only way documents are stored at CERN. If these web pages are reasonably structured and access to search engines is provided, both storage and retrieval seem to be perfect. Seen from the view point of LTEA, however, the web is a disaster:

- the structure of web pages changes frequently (as it is so easy to modify it),
- no "versioning" is done, i.e. when web pages are updated, the previous versions are not kept,
- no systematic control is done for "broken links", i.e. pages are deleted or renamed without a systematic attempt to update all the links pointing to them,
- although efforts are made to reduce the number of web servers at CERN, there are still a large number of "small" and "private" servers (many of them not even systematically known), with a very unsure lifetime: if the person responsible for such servers leaves CERN, the pages are not necessarily moved, and are lost. **As a result, the amount of information lost in this way by CERN is unknown.**

##### *Certified Information Systems (CISs)*

The two systems that were studied can be described as follows:

- they satisfy to a large extent the requirements of "real" archiving,
- as long as they are used, there is no danger in the medium term,

- they are not friendly to all authors or appropriate for all document types,
- they are not known or used enough.

#### *Databases*

Not studied.

#### *Paper*

This was included in this list because all paper documents are candidates for LTEA as well. Some of the classical archive rules can be applied to LTEA, but not without certain modifications, due to the electronic storage.

#### *Other, i.e. not belonging to any of the environments mentioned above*

This is the most worrying possibility, because very little of it is known, but we can be sure that valuable information is kept on local disks of desktop computers, in "private" servers, etc., with no links in the other environments indicating their existence.

### ***3.3.2 Description by document type, structure, complexity***

In some divisions, there exist very detailed lists describing which document types are used in the division; other divisions have no explicit knowledge about the material they are dealing with (only present in the minds of those responsible).

### ***3.3.3 People responsible for documents***

There is no easily identifiable such class of people, and only a generic list can be given as follows:

- authors of individual documents,
- everybody sending e-mail,
- secretaries (also of meetings),
- web masters and the people they are collaborating with,
- users of CDS, EDMS and similar documentation systems
- Divisional Records Officers (DROs).

## 4. Recommendations: Towards an affordable LTEA

On the way towards the implementation of an affordable LTEA at CERN we certainly have to follow the slogan:

Think globally,  
Act locally

With this in mind, the recommendations take into account the interrelatedness of the various aspects of LTEA in its general and CERN-specific form, but propose individual actions to be taken. The level of urgency and cost are determining factors in prioritizing the proposed actions.

The guiding principles are:

- to prevent further loss of information as soon as possible,
- to concentrate on a document-handling policy as a basis for LTEA,
- to postpone genuine LTEA where it can be done without risk.

Our recommendations first present the concrete actions to be taken now and the **preparatory measures** – to be taken in parallel – as a prerequisite for successful and yet affordable LTEA on the CERN scale; then follow the **accompanying measures** that will improve the usability and the "contextual information" of CERN's documents on a general level; finally those parts of LTEA that can be handled later are enumerated.

Apart from a few exceptions, very little information is given concerning the cost of LTEA implementation. On the one hand, some important actions would be handled by in-house working groups or task forces. On the other hand, it is considered premature to give any estimates for the LTEA, as long as a CERN-wide document-handling policy has not been defined and implemented.

### 4.1 Concrete actions

For purely practical reasons the recommendations concerning concrete actions are organized according to the different storage environments: e-mail, web, CIS, databases, paper and others.

#### *4.1.1 Actions to be taken in the area of electronic mail*

Following the principle of "preventing further loss", the WG obviously would have to propose archiving every single mail sent to or received by any person having a CERN mail address. This was felt – at least for the time being – to be exaggerated and expensive. The recommendations in this field are:

- *Reduce use of "private" mailing lists through publicity for listbox-lists archives:* people must be made aware of the fact that all mail sent to listbox lists is already being (pre)archived now, with the aim of convincing them not to use "private" mailing lists for anything that may be considered work-related. This could be done by directly addressing group leaders



and list owners and members, and by articles in the CNL, the Weekly Bulletin and other forms of publicity and, also, by improving existing documentation.

- *Introduce archiving of selected mails:* introduce archiving for a small number of important people (higher management, spokespersons of experiments and people designated by them). The cost of such an action would be almost zero, but it needs a certain degree of collaboration from the selected people (if their "working tools" change). The selected people would also be able to "opt out" for specific mail messages. The number of selected people would have to be kept small (of the order of 50) (see Annex 3).
- *Introduce archiving of mail folders:* offer – to every individual who opts for it – the possibility of archiving his/her mail folders. The cost of the implementation would be of the order of 50 to 70 kCHF and the work would represent one month by a trained technical student (see Annex 3).

#### **4.1.2 Actions to be taken in the area of normal web pages**

This denomination covers all those pages that are not stored in a particular document-handling system, even if they were produced by "web editing tools". Since no archiving efforts are made for normal web pages on the one hand, and in view of their enormous popularity on the other, we can neither leave the situation as it is nor force everybody off the web. In view of the cost and the technical difficulties involved, and despite the urgency, there are no fast solutions. We can therefore recommend a medium-term solution, which should follow two parallel paths:

- *Reduce "wild" usage of web:* move as many normal web pages as possible into document-handling systems. A forced migration would be considered contrary to the CERN culture (freedom). It is therefore unlikely that such a migration could be done quickly for a large part of these records. Also, even if all CERN staff could be forced to move their documents, a lot of the experiments-related material would stay outside, if the responsible persons are not convinced to use appropriate systems. However, a serious effort must be made to discourage authors from storing their new documents "somewhere" on the web.
- *Consider web archiving:* do not accept the offers so far received from companies for a feasibility study, but pursue the search for a web-archiving system implementable at CERN through other projects: Kulturarw3 of the National Library of Sweden, the Internet Archive (non-profit organization based in San Francisco), and the Minerva Prototype of the US Library of Congress.

A desirable side effect of web archiving would be the systematic check for broken links. The minimum that should be done is to continue to reduce the number of web servers, and eliminate "private" servers completely.

**NB:** The idea of archiving *all* CERN web pages goes clearly against the classical concept of archiving, in which the selective approach plays an important role. However, modern search engines (and even more so the ones that will be available in the future) make the problem of selective retrieval disappear.

→ **Pursuing web archiving could be done by B. Pollermann, if agreed officially.**

These recommendations obviously compete with each other: if one were applied 100%, the other would be unnecessary. However, following both paths increases the

security of records to a desired level, while still leaving some freedom to the authors of documents.

→ In any case the exact direction to follow, especially which of the normal web pages to move and into which system, has to be defined within the CERN-wide document-handling policy. Whichever way we go, it should be clear that all information stored only on the web will be lost sooner or later, unless web archiving is implemented at CERN.

#### *4.1.3 Actions to be taken in the area of CISs*

Certified Information Systems must be used as much as possible for storing documents at CERN. While more general recommendations will be given in the "CERN-wide documentation policy" section, two specific recommendations can be given here:

- *Minimize the number of CISs*: in the long run, only the two systems CDS and EDMS should be retained, unless users and authors provide reasons for keeping more systems good enough to justify the additional cost of their maintenance.
- *Continue to improve CDS and EDMS*: in order to increase their acceptability for users and their value for the LTEA, we should implement the missing points to fulfil all the requirements for a real CIS; adapt to new document types; increase the capacity.

#### *4.1.4 Actions to be taken in the area of databases*

Before any further recommendation can be made, long-term aspects of databases have to be studied.

→ Must not be neglected, but would need a special mandate.

#### *4.1.5 Actions to be taken for paper-based documents*

Documents stored on paper only should be at least registered into a CIS with title and meta-data. As long as the documents now stored on paper are not legally sensitive, they can also be stored exclusively in electronic form. However, further study must be undertaken before this can be done with legal documents.

→ Needs collaboration with the Legal Service.

#### *4.1.6 Actions to be taken for documents in "other areas"*

Here, the most urgent action is to gather information about the extent to which documents are stored outside the environments mentioned in the preceding paragraphs. Example: Communication concerning contracts could be done by e-mail to mailing lists.

→ Would have to follow from the document-handling policy. However, the management (on the level of group leaders), in collaboration with the DROs, must immediately begin work on identifying the cases.

### **4.2 Preparatory measures: A CERN-wide document-handling policy**

Rule 5 of the subsidiary document to the operational circular N° 3 (OC 3: Rules applicable to archival material and archiving at CERN) stated: "Archiving not only involves the preservation of documents of long-term interest, but also requires organised records management and the application of defined standards at the moment of creation

of the document". Defining and implementing an official document-handling policy to be followed CERN-wide would prepare the ground for successful, and yet affordable LTEA.

Implementation of an official document-handling policy would also

- free manpower on the system maintenance side by not spreading resources over several competing systems;
- save time for the authors (i.e. for those people who maintain documents) by answering their eternal question: "Where should I store and maintain my documents?";
- save the readers' time and reduce their frustration (which also costs manpower!) by facilitating the search for the desired documents and their access.

Thus, in times of shrinking resources, the above benefits constitute additional arguments in favour of a CERN-wide document-handling policy.

While these arguments may appear as unrelated to the question of LTEA, we believe that it would be less costly to implement LTEA when a well organized system is in place. Once all have agreed that CERN needs a document-handling policy, the requirements for LTEA must be taken into account when deciding on this policy. This consideration leads directly to the necessity of using CISs (see as Annex 2 the draft of a summary by Anita Hollier, the CERN archivist).

The general principle for a document-handling policy could therefore be formulated as follows:

**All CERN documents of any official importance  
must be stored and maintained  
in systems fulfilling the requirements of CISs.**

In addition the document-handling policy has to provide answers to the following questions:

- Which of the "near-CIS" in use at CERN (e.g. CDS and EDMS) should be retained and supported?
- To what extent should they be developed (increase of capacity, flexibility, adaptation to new types of documents, user requirements)?
- Which other forms of documentation storage would still be allowed or tolerated?
- To which degree would other systems (or "storage behaviour") be scaled down or stopped altogether?
- If more than one CIS is chosen: Which CIS is most appropriate for which type of document? (Minutes, for instance, can – to the distress of authors and readers alike – be found in all storage environments, except databases.)
- How, by whom, and over what time would the migration from undesirable storage environments and behaviour towards the CIS be handled?
- In which way could the participants be motivated (although this is still easier than for genuine LTEA)?

- Which additional resources can be assigned to the implementation of the document-handling policy in all its aspects, given the fact that in the long run CERN will save resources?

**NB:** The term "document" is being used here in a very wide sense. It ranges from a simple note, agenda or minutes, over newsletters, to very complex technical documents with pictures or photos, etc. In this sense it is even more general than the definition of "Archival material" in the OC 3.

The creation of a task force

The task of defining a CERN-wide document-handling policy and ensuring its implementation is urgent for LTEA, and desirable in general for the reasons given above. It needs to be handled by a relatively small but dedicated task force (3 or 4 people) that should be set up quickly. Here are some points for the task force to consider.

The first step would be to collect information that could serve as a basis for further steps. This would include

- a list of all types of documents produced or maintained in the different divisions, accompanied by information on how and in which system the documents are stored;
- a list of the people or units mainly responsible for the creation of these documents;
- a list of representative users (readers).

→ **Part of this task will be done in a meeting of DROs.** If encouraged, a systematic assessment will be possible.

The next step would be to build a list of user requirements.

→ **This step was partially realized by AS (for the Desktop Forum).**

The results would be handed over to the task force for analysis with the aim of "distilling" from these raw data the justified and defensible requirements for a document-handling system. These requirements would then have to be merged with the requirements of a CIS.

Armed with these requirements, the task force would then invite the persons responsible for major document systems (certainly EDMS and CDS, but maybe more if there is strong interest after a CERN-wide announcement) to defend their own, and prove that it fulfils all the justified and defensible requirements. Representative users would be heard on the aspect of user friendliness.

On the basis of the results of this "contest" the task force could finally decide which CISs are to be retained and which of their missing features are to be implemented (after finding out their cost).

What remains to be done is to give an answer to the outstanding questions. Once the successful implementation of the policy is at the horizon, the LTEA aspects will be followed up again, and implemented (possibly by another task force launched by the Archive Committee).

A possible decision to purchase a completely new system from outside should not be discarded from the start. In this case, however, the LTEA features must already be present in the product itself.

### 4.3 Accompanying measures: foster meta-data

Meta-data are a delicate issue. The busy author, eager to get his document published, considers them as a source of additional work without obvious recompense. The individual reader might not even know about the concept of meta-data. However, not only the people who look at documentation as a whole (e.g. archivists), but also users who try to find certain information, appreciate and profit from high-quality meta-data. For these, they are an inseparable part of the records.

The following compromise between these opposing positions is recommended:

- *Offer good guidelines* (a standard): given the relative simplicity and wide distribution, the DC (Dublin Core) should be applied.
- *Define a minimum required for all documentation*: the absolute minimum should be a meaningful title (also called Subject in the area of e-mail) and an abstract. For small documents and e-mail, an abstract for a set of documents or a list would be sufficient. Also, the name of the document should follow some reasonable rules (see next section).
- *Help authors directly*: giving some of their work to "moderators" will help authors in fulfilling the minimum requirements and also assure a good quality. People not directly involved in the production of the document, but with some experience in this task, are often in a better position to create good meta-data.
- *Rely on the excellent performance of modern search engines for the rest*: if such an engine (e.g. Google) has access to the full text of the documents, it is in many cases not necessary (or can even be harmful, if not done professionally) to specify too many keywords in the meta-data.
- *Give documents a good name*: Document names must facilitate finding a desired document amongst many others. This can be achieved by giving names that are easy to remember, easy to guess, and easy to spot in a lengthy list generated by a search. The less arbitrary and the more meaningful the names are, the better they will fulfil this role. Furthermore, they must be as simple as possible.

This last feature can be controlled by naming conventions applied to the URLs of documents, above all to those stored in a CIS and accessible through the web. The following naming conventions describe the structure, the content and the form. Thus in choosing names the authors should make sure that they

- consist of meaningful English words or known acronyms (defined in a special collection of acronyms),
- have a hierarchical structure with the higher-level parts starting from the left,
- have clearly defined separators between different levels,
- consist of lower-case letters only.

An additional rule should be applied for names within document structures: the structure must be such that the higher levels always really exist, i.e. that they can be accessed through the web.

Example: Let us assume that there is a working group with the name 'eawg' and that its minutes are stored in a CIS.

A good name for this document would be

<http://cern.ch/minutes/eawg>

and it would be mandatory that there also exists a document

<http://cern.ch/minutes>

This page – maintained automatically – would give access to all minutes of all meetings at CERN.

**NB:** In many cases, the hierarchical order is not unique for one document. If – in our example – the 'eawg' people decide to have their own web site with the name <http://cern.ch/wgs/eawg>, their minutes would become a subset of this URL and be found under <http://cern.ch/wgs/eawg/minutes>. Such use of aliases is also highly recommended as it favours the guessability of names.

These rules should be applied to the names of mailing lists and the names of web pages, while list names related to them should also contain the same name (apart from the separators between name parts, which have to be different for technical reasons). For example, in the case of a project called xyz, the web site name would be <http://cern.ch/project/whynotxyz>. The name of the mailing list would then be [project-whynotxyz@cern.ch](mailto:project-whynotxyz@cern.ch), since the / is not allowed in names of mailing lists.

→ **Naming conventions in this sense already exist and names are being moderated** (by IT). It is, however, recommended that naming conventions be improved still, and that a more official status be given to the moderation activities.

#### **4.4 Actions and decisions to be postponed**

In the previous sections, the recommendations concerned pre-archiving, i.e. making sure that:

1. all potentially valuable records are stored and can be retrieved without any loss over a short- or medium-term period (e.g. 1–5 years),
2. nothing is done that could harm the implementation of genuine long-term archiving (30 years).

This section is about the actions and decisions related to genuine long-term archiving that can be postponed. The reasons that led us to consider the postponement of the long-term part are summed up below.

##### **4.4.1 General reasons**

Only recently have people become aware that valuable information may get lost forever. As a consequence, they realized that the state of the art in genuine long-term archiving is not yet satisfactory. However, interest is growing and more standard solutions may emerge. Prices will go down, so by waiting one will save money.

##### **4.4.2 CERN-specific reasons**

A CERN-wide document-handling policy is considered to be a prerequisite for LTEA, so CERN is not ready yet.

The effort that has to go into pre-archiving is considerable. Postponing certain actions allows us to concentrate on the urgent tasks. Data security (within the official systems) is sufficient, so there is no unacceptable danger of loss in these systems.

Postponing does not mean forgetting about these tasks. On the contrary, it is essential that the situation in archiving be monitored at regular intervals. What follows is a list of those areas in which actions and decisions can be postponed, but must be monitored:

- *Original format vs. converted*: to serve as a legal proof, a document has to be archived in exactly the same form as it was created. To be readable in the long term a document has to be converted into "tagged ASCII" (e.g. XML). Until these conflicting demands can be satisfied, it is recommended to keep both, the converted as well as the original version.
- **Regular checks should be made of the readability of the original format.**
- *Storage of pictures, drawings, photos*: the question of storing pictures etc. is part of the problem of formats. However, there is no obvious conversion (as XML).
- **Regular checks should be made of the readability of pictures, and of improved storage formats.**
- *Long-term vs. "normal" storage*: long-term archiving in the true sense implies a different way of storing and retrieving (bigger capacity, but slower), a regular change of the storage medium (to avoid loss via ageing) and recovery from disasters.
- **Regular checks should be made of the ageing of disks.**
- *Bulk vs. selective archiving*: the essence of classical (paper) archiving is the selection that is made before a record is sent to the archives. In the field of electronic archiving, bulk archiving is a meaningful possibility since the advent of cheap mass storage and retrieval by search engines.

The problem that will remain in any event with selective archiving is simply that nobody knows what will be important for the historian in the distant future.

Example: The first web pages of a student who will turn out to be a Nobel prize winner 30 years later will always fail rigid criteria. Also, according to the US Library of Congress, selective archiving is 100 times more expensive than bulk archiving.

→ **No regular checks are needed, but a political decision must be taken.** In practice the monitoring could be initiated and supervised by the Archive Committee, provided they are given the necessary resources.

#### 4.5 Authors motivation

Motivation is always important, and it is especially difficult to maintain in such an abstract and elusive area as LTEA.

Motivation is high for tasks directly relevant to the single person involved, and if they concern her/his immediate future. LTEA characteristics are exactly the opposite: the benefits are for CERN as a whole, going over very long periods of time, most likely

beyond a person's stay at CERN, and probably beyond her/his own lifetime. The average author will therefore not feel concerned, and will not be motivated to contribute to LTEA.

For these authors, motivation must be "artificially stimulated", first by making it easy to comply with the demands of LTEA, then by a slight pressure being applied from the management side.

LTEA must be made easy for authors by:

- running the archiving part completely automatically (in the background) after the author has "published" (i.e. submitted the document to the recommended system),
- having clear rules of how to publish, and making them known,
- offering features appropriate for each type of document, and
- keeping the work needed for publishing negligible with respect to the work for creating the document itself.

Once these conditions are fulfilled, the author can feel further encouraged if it is clear to him that a document not published according to the rules will not be taken into account for the MOAS achievements.

Finally a special campaign must be launched to reach the critical mass for authors using the recommended method, so that further motivation is achieved by peer pressure.



## **5. Summary of Recommendations**

After analysing the problem of LTEA in general and assessing the situation specific to CERN in this field, the following recommendations can be made:

- implement the e-mail archiving for selected users and folders,
- pursue the web archiving with non-industrial institutions,
- define and implement a CERN-wide document-handling policy,
- postpone genuine long-term archiving, as long as no information is lost, in expectation of cheaper solutions,
- be attentive to the possibility of loss of access due to format changes and similar.

## 6. Conclusion

Since the nature of the mandate was clearly action-oriented, this report presents a programme of work rather than a set of final recommendations. It also highlights the fact that the transition from the present state to a fully functional LTEA is a process to be developed over a length of time that will depend on the available resources.

The biggest problem – for which there is no quick solution – turns out to be the enormous popularity of the web as a document storage, due to the ease with which it can be used.

Provided it is given the necessary resources, the Archive Committee can initiate and supervise all tasks directly concerned with LTEA. However, a task force needs to be set up beforehand and quickly to define and implement a CERN-wide document-handling policy.

Bernd Pollerman  
on behalf of the LTEA Working Group

Endorsed by Gabriele Veneziano  
on behalf of the Archive Committee

**List of the Members of the  
Working Group on Long-term Electronic Archiving (LTEA)**

Sergio Cittolin, EP  
Vittorio Frigo, AC  
Marco Ganz, IT  
Mike Gerard, IT  
Michel Goossens, IT  
Anita Hollier, ETT  
Jean-Yves Le Meur, ETT  
Monica Marinucci Lopez, IT  
Mats Moller, AS  
Corrado Pettenati, ETT  
Thomas Pettersson, EST  
Bernd Pollermann, IT (Chairman)

## CERN Certified Information Systems (CIS)

### Contents

Introduction

System requirements - Synopsis

System requirements - Detail

Metadata requirements

### Introduction

The primary requirements of CERN Certified Information Systems are that they

- facilitate good electronic record keeping, in order to improve information management and operational efficiency
- manage the records in such a way that selected ones can be preserved permanently as valid archival records<sup>1</sup> (either within the CIS itself, or in a central electronic Archive).

This document does not address all the features required in good electronic document management systems (EDMS) or electronic records management systems (ERMS); broader guidelines on this are available elsewhere.<sup>2</sup> Instead it focuses on the second point: elements essential as preparation for the long-term preservation of records.

### System requirements - Synopsis

1. The system must preserve all the aspects of a genuine record:

---

<sup>1</sup> The International Council on Archives defines a record as " ...a specific piece of recorded information generated, collected or received in the initiation, conduct or completion of an activity, and that comprises sufficient content, context and structure to provide proof or evidence of that activity." (Conseil international des archives - International Council on Archives. Guide For Managing Electronic Records From an Archival Perspective. Committee on Electronic Records, February 1997. (ICA

Studies/Études CIA 8) Section 2.1)

<sup>2</sup> See, for example:

US Department of Defense standard DOD 5015.2-Std (Design Criteria Standard For Electronic Records Management Software Applications).

<http://jitc.fhu.disa.mil/recmgt/index.htm>

Model Requirements for the Management of Electronic Records (MoReq)

<http://www.cornwell.co.uk/moreq.html>

UK Public Record Office - Functional Requirements and Testing of Electronic Records in Management Systems

<http://www.pro.gov.uk/recordsmanagement/eros/invest/default.htm>

Content: what the record "says", i.e. the actual data object

Structure: the internal format and arrangement of the record (which may comprise more than one document)

Context: information on where, when, why and by whom the record was created and its role in the activities of CERN.

2. The system must preserve the integrity of the records; this includes:

Version control: the identification of revised records and safeguarding of obsolete records

Inviolability: the prevention of unauthorised access, alteration or removal of records

Authenticity: preservation of a history of the record's creation, transmission and use.

3. The system must preserve long-term access to the records.

In order to encourage its use, a CIS should offer added value to authors by allowing them to manage their information more efficiently. For this reason the systems should be user-friendly and should be able to manage *all* the records of the user group, not just those intended for permanent preservation. The CIS should capture, or prompt the author to provide, certain minimum metadata for all records when they are filed. Additional metadata may be added at the author's discretion.

### **System requirements - Detail**

This list gives some key requirements, but is not exhaustive.

Mandatory requirements are in plain type; less crucial ones are shown in *italics*.

1. The system must preserve all the aspects of a genuine record

- The CIS must capture, or prompt the author to provide, certain minimum metadata for all records when they are filed (see metadata list below). Additional metadata may be added as required. Thereafter, only authorized individuals may change author-supplied metadata, and metadata captured by the system may not be altered at all.
- The CIS must link electronic record content and metadata together in a tightly bound relationship.
- *The CIS should provide the capability to output metadata for viewing or printing.*

2. The system must preserve the integrity of the records

- The CIS must support the creation and maintenance of a classification scheme or fileplan to which all electronic records will be classified.
- The reference codes assigned under this fileplan should be unique; if they are not, the CIS must assign a unique computer-generated record identifier to each record.
- The CIS must provide the capability to store version(s) of a record and to link original superseded records to their successor records.
- *When the user selects a record for retrieval, the CIS should check for the latest version of the record, but allow the user the flexibility to retrieve any version.*
- The CIS must *either* safeguard all obsolete versions of every record, *or* it must allow working documents to be 'declared' as records, after which no further revisions are allowed. Revised versions of these 'declared' records will then automatically be saved as a new version, and obsolete versions will be safeguarded.
- The CIS must provide the capability to define different groups of users and access, and to prevent unauthorized access to records.

- The CIS must preserve a history of the record's creation, transmission and use, and any alterations made to it (usage log).
3. The system must preserve long-term access to the records
- The CIS must be supported by procedures that ensure migration of records and maintenance of media as required in order to ensure that records remain accessible.
  - Apart from these authorised migrations, the CIS must maintain the integrity of the record as it was created / received, and must not allow any alteration to it.
  - The CIS must enable electronic records that are selected for permanent preservation to be exported to an electronic Archive without loss of either content or metadata information.
  - The CIS should allow access to an ASCII / XML version.
  - The CIS must provide the capability for authorized users to tag a record for permanent retention. *It should also allow the addition of more detailed information about retention periods based on an authorised disposal schedule)*

#### Metadata requirements

A simplified metadata list is given below, based on Dublin Core<sup>3</sup>. Mandatory requirements are in plain type; less crucial ones are shown in *italics*.

The CIS should be able to capture some of this information automatically; it should prompt the record author/owner to provide the rest.

- Title - The name given to the resource by the creator or publisher
- Creator - The person or organisation primarily responsible for the intellectual content of the resource, e.g. author of a document. Maximum provenance information should be given, including name, affiliation and role (e.g. position within CERN, or other institution).
- Subject - The topic of the resource. This should preferably be based on key words and controlled vocabularies, and may include reference codes from the classification scheme by which the resources are arranged.
- Description - A textual description of the content of the resource, e.g. abstracts of documents. This should also include additional descriptive information about the activity / transaction represented by the record.
- *Publisher - The entity responsible for making the resource available in its present form.*
- *Contributors - Persons or organisations, in addition to those specified in the creator element, responsible for making significant intellectual contributions to the resource.*
- Date - Date of creation of the resource (ISO 8601 (W3CDTF) recommends the YYYY-MM-DD format).
- Type - The nature or genre of the content of the resource, preferably selected from a controlled vocabulary.
- Format - The physical or digital manifestation of the resource. The intention is to give sufficient information to allow people or machines to make decisions about the usability of the encoded data, e.g. what hardware and software might be required to read it.
- Identifier - An unambiguous reference to the resource within a given context, e.g. the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)) for networked

---

<sup>3</sup> Dublin Core Metadata Initiative

<http://dublincore.org/documents/dces/>

resources, or a unique reference code from the classification scheme by which the resources are arranged.

- *Source* - The work from which the resource was derived, if applicable.
- *Language* - The language(s) of the intellectual content of the resource, preferably following ISO notation.
- *Relation* - Relationship to other resources.
- *Coverage* - The extent or scope of the content of the resource. Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
- *Rights* - Information about rights held in and over the resource; conditions of access and use.

A. Hollier ETT/SI  
21 June 2001

## Email Long Term Archiving

In a world in which electronic documents can be produced and exchanged with unprecedented easiness a new paradigm for archiving is needed.

Specific electronic archiving facilities are necessary to protect CERN from the risk of loss of data that could become critical in the long term.

One of the domains in which CERN is more exposed to this risk is Electronic Mail (email), currently the technology most used to communicate by the members of the Organisation.

Two fundamental needs must be satisfied by an electronic archiving facility for email:

- archiving of all messages for selected users;
- archiving of selected messages for all users.

The EArchiving Working Group (EAWG) stresses the need for evolution of the current Mail Service to satisfy these two requirements, first step in the direction of making the service become a real CIS.

Archiving all messages for selected users

=====  
Critical decisions are taken nowadays which are increasingly based on email exchanges between the CERN Directorate and other institutions.

In order to keep the possibility of understanding these decisions in future, the whole mail traffic of the CERN Directorate should be kept forever.

Technically this could be easily implemented in the current system by combining the usage of listbox archives and mail forwarding. Users would keep the possibility of excluding messages which shouldn't be archived (opt-out approach). Filtering techniques could be used to make the exclusion automatic in specific cases (e.g. personal messages).

1. Two mailing lists should be created for each selected user, e.g.:

earchive-maiani-incoming and earchive-maiani-outgoing

This lists should have no users, be completely closed and archived. Incoming and outgoing traffic could be merged into one single list too though keeping the two flows separated might be cleaner.

2. An automatic forward should be set on the user's account so that a copy of each incoming message would be sent to the corresponding list (earchive-maiani-incoming). A filtering software (e.g. procmail) on the server could be used to automatically exclude specific (personal) messages from the archive.



3. The user's mail clients must be set to 'BCC' automatically the outgoing list (earchive-maiani-outgoing) so that each outgoing message is saved unless the 'bcc' is explicitly removed by the user (which would therefore keep control over what's archived).

#### Archiving selected messages for all users

=====

The possibility to store messages into Folders on the Mail Server is the first step towards moving the documents into a highly reliable centralised system. That's not sufficient anyway since:

- quotas are imposed on the folder space that force sometimes users to move messages out of the Mail Server;
- users maintain read/write access to the folders while this possibility should be restricted to the archivists.

A reliable write-once/read-only repository with virtually unlimited size restrictions should be provided. Technically this could be implemented in different ways, depending on the Mail Server architecture to be deployed.

Three possible solutions:

#### 1. Proprietary archiving system

Archiving solutions are available on the market which are well integrated with the most common commercial mail/groupware products.

Advantages:

- complete integration with the server backbone;
- archiving options and features are added to the mail clients to make the distinction between the mail server and the archiving system completely transparent;
- possibility to store and link different kind of objects (not just mail);
- commercial support;

Disadvantages:

- requires specific commercial mail servers and clients (Microsoft Outlook/Exchange, Lotus Notes);
- long term availability;
- closed architecture;
- price;

#### 2. Dedicated IMAP server

A dedicated IMAP server could be setup to store user's folders for long term preservation without any quota, with read-only access and with the possibility to define access control lists (i.e. to make folders visible by several selected users).

Advantages:

- open architecture like the current mail system;
- blackbox visible by any IMAP client;
- all users would become aware of earchiving (maiani.earchive.cern.ch would appear close to maiani.mailbox.cern.ch);

Disadvantages:

- home-made solution;
- mail specific (how to make searches?);
- folder sharing can be tricky;
- user support;

### 3. Web access to the folders

Selected mail folders could be converted into web pages stored into a dedicated Web server.

Advantages:

- cheap;
- completely open;
- long term preservation (HTML files!);
- powerful search tools and authentication/authorisation mechanisms (inherited from the web);

Disadvantages:

- lack of integration with mail servers/clients;
- scalability?
- need to build all interfaces for submission/management;
- accessibility (needs its own interfaces);

The EAWG stresses the need to make long term archiving of selected folders one of the requirements to be satisfied.

M. Ganz  
23 Aug 2001