

Hi to all,

Following Tom's suggestion "I do believe we need to design this all 'on paper' before beginning to implement", I'll go through your last email (attached below for your convenience) and make some comments.

Let me remind you that in the CERN definition there is no one-to-one direct mapping between "document types" and "collections". They can be different. Primary (or real) and specific (or virtual) collections are mainly defined by the tag identifier or a query respectively. Instead, "document type" is used for the submission of the document and to access the full-text. We are now discussing about document types. These will appear in the main submit page, not in the main search page.

- > Even if these types do not cover any document one
- > could think of, it has proven to be helpful not to offer too many
- > different types. Most of them will not or rarely be used.

I think we all agree on this. We should choose the minimum number of document types necessary to insure adequate document's submitting.

- > We do not enforce naming conventions in addition to the type.

While this is feasible for engineering data or documents, we should keep in mind that physicists like to have a "meaningful" name associated to a document, especially an internal note or article. Personally, I particularly like the "year" field in the name. Ideally, I would retain both (the numbering and the naming). It is also true that what is really important is to guarantee the search and retrieval of records (by a wise definition of metadata).

- > Again, the number of attributes
- > has to be limited in order to ensure their utilization. Retrieval is
- > then based on any combination of attributes. [...]
- > Depending on the document type we offer a subset of matching attributes
- > to describe a document. The rest of possible attributes is hidden from
- > the user. Some of these type specific attributes are required ("r" in
- > the table), some of them are optional ("o").

If I understand, what you call "attributes" are the metadata associated to the document. I agree that their number should be the minimum necessary to insure and facilitate search and retrieval. In the CERN scheme, you can associate also "keywords" to record's metadata. The user could choose them by a predefined list, or input them by hand (free text) [as we discussed last week]. As soon as we have a list of document types, we will have to define the metadata we want to collect [see Tom's email about metadata].

Putting together all your suggestions (Tom+DESY), there are many common points. For the moment, let's concentrate on the document types we would need to define in the Invenio system (ILCDOC). ILCDOC should collect all the documents that are not in Invenio (talks and meetings) and that are not in the EDMS (specifications, technical drawings, quality management, expertise-?, ...).

[ I don't know exactly what kind of documents we would need to store for a given engineering project. Tom, could explain this to me or give me a better idea of what we

plan to store in EDMS? That's probably an oversimplification, but I guess that we would need something like:

- Ideas ←
- Technical specifications
- Drawings
- Assembly procedures or instructions
- Travellers or construction documentation
- Quality control data
- Test results ←
- + Construction schedule
- + Construction costs

The arrows indicate items that could be in the form of a note, and that could be submitted to ILCDOC, let's say "physicist-like" documents.]

"Files", "notes", "letters", "meeting minutes", ... : all of them are "Internal Documents". This could be view as just one document type. Personally, I would prefer splitting it in "Note" and "Miscellanea" (or "General Documents" if you prefer), because of the usual massive production of notes by physicists. What about:

### ILC Internal Documents

- Internal Notes ( NOTE\_\* ) ← CATALOG  
← DOCUMENT TYPE
  - ILC Area System
    - a. e-Source
    - b. e+Source
    - c. DR
    - d. RTML
    - e. Linac  
NOTE\_LINAC\_yyyy\_nnn or  
LINAC\_NOTE\_yyyy\_nnn
    - f. BDS
  - ILC Detector
    - a. SID ← CATEGORY  
NOTE\_SID\_yyyy\_nnn or  
SID\_NOTE\_yyyy\_nnn
    - b. LCD
    - c. GLC
  - ...
- Miscellanea (XLS files, tables, short DOC/TXT files like minutes or memos, letters ...) ← DOCUMENT TYPE
  - Correspondence ( CORR\_\* )  
(Letters/Email/...)
  - Minute ( MIN\_\* )
  - Memo ( MEMO\_\* )
  - Spreadsheet ( XLS\_\* )
  - Other

In this way we end up defining 2 new document types: “Internal notes” and “Miscellanea”. Each document type can have different categories, in such a way that the submit pages and the names of the records can be different.

Note that:

- We can add categories to the “Internal Notes” document type. This is not going to affect the number of document types. Categories are similar to what you call “subtypes” ( “implemented to further differentiate between the described document types”).
- The category can be added to the record name (ex. SID), or can be automatically stored as a keyword.
- All the Letters/Minutes/... are stored with the name CORR\_yyyy\_nnn/MIN\_\*/..., independently on the subject of the document and on the sender (committees or individuals). Do we like that?
  - In order to retrieve a specific letter/minutes/memo/..., the user can search for it in a Virtual Collection (ex. institution,...), provided that the keywords have been assigned correctly by the submitter (we can provide a list of predefined keywords to choose from, one list for each category).
- We have to associate these documents to the appropriate Real Collections (ex. “Publicatons&Notes” and “Communications&Agreements”)
- This is slightly different from Tom’s proposal, in particular
  - He suggests to have a “Files” document type, with categories “ILC project files” and “ILC Detector files” (implicit differentiation based on the source of the document)
  - He suggests to have a special “Letters&Minutes” document type

To stick to the CERN example (<http://documents.cern.ch/EDS/current/access/index.php>) we could also have:

<b>ILC Library Catalog</b>		← CATALOG
- <a href="#">ILC Reports</a> ( ILC_yyyy_nnn )		← DOCUMENT TYPE
- <a href="#">...</a>		
<b>Other</b>		
- <a href="#">ILC Administrative Documents</a>		

Note that:

- The Reports document type could have categories (BCD, RDR, TDR), mainly for the naming purpose, that is to have the “BCD”, ..., string in the record name.
- Which kind of “Administrative Documents” we will have? What do you store in it at Desy ? Depending on the answer, this document type could be redundant or not.

To resume, we could define the following catalogs /document types / categories:

<b>ILC Internal Documents</b>	←CATALOG
- <u>Internal Notes</u>	← DOCUMENT TYPE
• ILC Area System	
a. ...	
b. ...	
• ILC Detector	
a. ...	
b. ...	
• ...	
- Miscellanea	← DOCUMENT TYPE
• <u>Correspondence</u>	
• <u>Minute</u>	
• <u>Memo</u>	
• <u>Spreadsheet</u>	
• <u>Other</u>	
<b>ILC Library Catalogue</b>	←CATALOG
- <u>ILC Reports</u>	← DOCUMENT TYPE
- <u>External Notes/Publications</u>	
<b>Other</b>	
- <u>ILC Administrative Documents</u>	
- ...	

From ▶ "Markiewicz, Thomas Walter" <twmark@slac.stanford.edu>  
Sent Friday, September 22, 2006 5:59 pm  
To maura@fnal.gov  
Cc Silke Eucker <silke.eucker@desy.de> , Jasper Dammann <jasper.dammann@desy.de>  
Subject RE: CDS / EDMS configuration  
Attachments ILC tree for \_\_\_\_\_CDS-v1.doc 35K

Hi Maura,

There is so much in CDS I still find it hard to see the nuances of its organization. After looking through your earlier emails, maybe the way to start to design a system for ILC is thinking about ILC equivalent for the CERN submit page (as opposed to the search page)

<http://documents.cern.ch/EDS/current/access/index.php>

Files (Any openly shared unreviewed file of any type)

- ILC project files
- ILC Detector Files

Internal Notes

- ILC Area System Notes
- ILC Detector Notes

Letters& Minutes

- Committees
- Individuals

Media

- ??

ILC Reports

- BCD
- RDR
- TDR

I attach the file I promised to send yesterday. There is not much added content, just the beginnings of collecting what metadata we want for each document type.

I do believe we need to design this all 'on paper' before beginning to implement.

Still trying to find the time to do this with people here at SLAC; perhaps with the recently offered Fermilab help, critical mass will be reached & we'll make some progress.

Tom

From ▶ Silke Eucker <silke.eucker@desy.de>  
Sent Wednesday, September 27, 2006 7:56 am  
To maura@fnal.gov , twmark@slac.stanford.edu  
Cc Kathrin Lappe <kathrin.lappe@desy.de> , Lars Hagge <lars.hagge@desy.de> ,  
Jasper Dammann <jasper.dammann@desy.de>  
Subject Discussion on ILC document types  
Attachments 060927\_desy\_documenttypes\_overview.xls 24K

Hello Maura, hello Tom,

with regard to the ongoing discussion about document types, we would like to add our experience.

Enclosed you find an overview of the document types in operation for some DESY projects. Even if these types do not cover any document one could think of, it has proven to be helpful not to offer too many different types. Most of them will not or rarely be used.

We do not enforce naming conventions in addition to the type. Instead, attributes which could be useful for retrieval are used. Part of the attributes are automatically handled by the system, the other part has to be entered by the document creator. Again, the number of attributes has to be limited in order to ensure their utilization. Retrieval is then based on any combination of attributes.

Depending on the document type we offer a subset of matching attributes to describe a document. The rest of possible attributes is hidden from the user. Some of these type specific attributes are required ("r" in the table), some of them are optional ("o"). Thus a minimum of common description can be ensured and it is possible to query e.g. for drawings with a certain paper size.

If considered necessary, additional subtypes are implemented to further differentiate between the described document types. These subtypes do not have any impact on the attributes. They just contain an additional typification which is used for e.g. providing different template files for various meeting series.

We hope this experience contributes to a fruitful discussion.

Best regards,  
Silke and Kathrin

-----  
Deutsches Elektronen-Synchrotron (DESY)  
Silke Eucker  
Informationsmanagement, Prozesse und Projekte (IPP)  
Notkestrasse 85  
22607 Hamburg (GERMANY)

Tel.: +49-40-8998-1772 E-Mail: silke.eucker@desy.de  
www: www.desy.de  
-----

