A case report of first steps taken at an European scientific agency,

European **C**entre for **D**isease Prevention and **C**ontrol

to introduce knowledge based services

László Balkányi
ECDC Scientific Advice Unit

MEDINFO2007
Brisbane, Australia
August, 2007

Abstract

This paper reports author's work toward building knowledge based services at the new European Centre for Disease Prevention and Control (ECDC). As part of this work, electronic content was analysed and metadata were extracted from a large document base.

ECDC, a scientific agency has been created to strengthen Europe's defenses against infectious diseases, aims to pool health knowledge, to develop authoritative scientific opinions about the risks. A quickly accumulating body of ~ thirty thousand electronic documents was analyzed.

As a way out from the current start-up phase, it was found that a set of systematic measures among them introduction of metadata enhanced document templates (using Dublin Core Metadata Initiative) and ontology based tagging should be applied.

As an innovative approach we used the existing arbitrary file and folder structure to be translated to content characterizing metadata. We hope that the in the next future we will be able to build up a knowledge repository with enhanced navigation facilities, using the existing electronic content – but made it much more transparent and easy to use source of knowledge for internal and external users at ECDC.

Introduction

ECDC[1] is a **scientific agency,** strengthening Europe's defence against infectious diseases.

The agency has four technical units (**Scientific Advice - SAU**, **Preparedness and Response - PRU**, **Surveillance - SUN** and **Health Communication – HCU - units**), an administrative unit and seven horizontal disease-specific projects[2]:

1) influenza,
2) tuberculosis,
3) food- and water-borne diseases,
4) other diseases of environmental and zoonotic origin,
5) vaccine preventable diseases and invasive bacterial infections,
6) HIV-STI and blood-borne viruses, and
7) antimicrobial resistance and healthcare associated infections.

Knowledge-based services (KBS) for human- and for machine users will be a key issue.

The *planned knowledge based services* are:

(a) electronic information sources service (content management)
(b) terminology services;
(c) other advanced systems / functions as knowledge navigation, decision support

In the area of knowledge management ECDC wants to avoid 'reinventing the wheel' and to stick to standards[3] ensuring semantic interoperability and transparency.

[1] ECDC will have a >50 M€ budget by 2010, staff will be ~ 300.
[2] At the time of the 'snapshot' for document analysis three units, SAU, SU and PRU were active, HECO was established later.
[3] on several levels: as the XML family, UML, OWL, DCMI etc.)

Materials and Methods

Our **materials** were the electronic documentation, accumulated since May 2005. Documents are stored on file servers, a common space serves shared documents, while virtual drives were assigned to relevant organizational units. A ‘snapshot’ of this structure was taken in August, 2006.

Files were classified into documents[1], spreadsheets, presentations, and ‘all others’ categories. **Methods:** Command line tools and disk cataloging utilities for extraction of file / folder structures were used; a text editor served for building/editing xml files, a spreadsheet application for file statistics.

[1] (working) docs, and (to be published) pdfs

Results

Characteristics of file accumulation and distribution were analyzed. Numbers in **table 1** summarize just the most characteristic aspects.
The folder structure is the result of a non-regulated, ‘organic’ growth in a just formed organization, where highly autonomous and highly educated experts worked mostly either alone or in ever changing, problem oriented, ad-hoc teams on emerging problems and most experts did extensive multi-tasking.
However **the amount of accumulating documents was / is staggering[1] (fig1.)**
The distribution of files is detailed according to the most frequently used file types in **table 2. Trends were calculated concerning file numbers (fig1)**.

[1] be aware that the value axis of fig. 1 is logarithmic!

Table 1

The Number of Folders / Subfolders

Subfolder depth level:	No. of folders in			
	"shared" drive	PRU drive	SAU drive	SUN drive
root folder	14	30	41	42
2nd	44	156	84	187
3rd	177	240	152	276
4th	384	183	68	142
5th	661	178	24	81
6th	383	91	9	42
7th	245	12	12	39
8th	194	8	1	
9th	188	1		
10th	109			
11th	80			
12th	27			
13th	9			
14th	1			
Sum of folders	2516	899	391	809
Total folders:	4615			

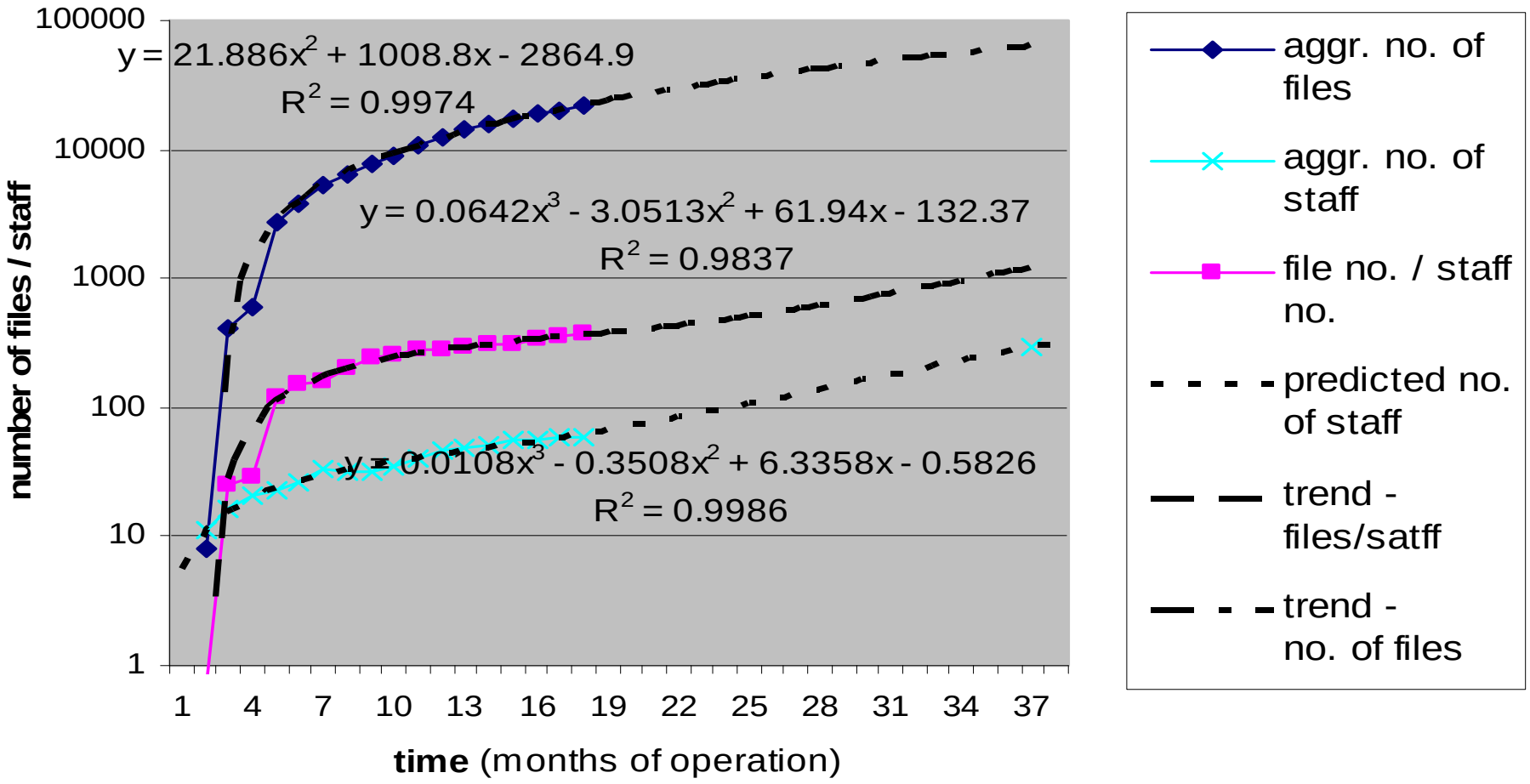
Table 2

File Type Distribution

location of files		Documents					
			Spreadsheets				
				docs to publish (pdf-s)			
					presentations		
						miscellaneous	
							all
shared 'drive'		32%	3%	22%	6%	37%	18264
PRU 'drive'		34%	4%	12%	5%	45%	7460
SAU 'drive'		25%	22 %	5%	2%	47%	3477
SUN 'drive'		40%	12 %	7%	3%	38%	5047
Tot.	abs. no	9185	1643	5108	1615	11650	29201
	rel. no	31%	6%	17%	6%	40%	

László Balkányi
ECDC Scientific Advice Unit

Figure 1 - **Trends of the Growth of Staff and the Aggregated Numbers of Scientific Document Files on ECDC Servers**



Discussion

Discussing files / folders structure:

It is almost hopeless to navigate efficiently in such a complex folder hierarchy tree. (Table 1)

The pan of short time memory (while working in folder hierarchies), allows us to handle 7(+-2) “chunks” of information at any given time [1]. The material, we deal with, however does not allow such a grouping – we have to opt for other organization methods as e.g. multidimensional localization by metadata.

Concerning search and retrieval:

Having thousands of documents on very similar topics, a conventional search engine will be never specific enough. Also, such a search will not find sets of (conceptually) related documents. Using metadata is of course the answer to this problem as well [2].

The metadata ‘vocabulary’ should be based on ontology, so that later on we will have the option of semantic operations as well.

Discussion – continued ….

Concerning the growth of the information heap:

ECDC will grow to size of ~ 300 staff in two years.

Well fitting trend lines (fig. 1.) show that the **file numbers will reach nearly a hundred thousand** and each staff member will be ‘responsible’ for more than a thousand file.

These predictions might be disputed even to be low-key, as more and more ‘services’ of ECDC will become ‘operational’.

Distribution of files types:

Table 2 shows us, that while the single most frequent information object is the ‘document’ file (31% of all), there is a huge heterogeneous set, containing raw txt, html, gif, jpg, rtf, eml, etc. and some others.

The difficulties to reach, to make available the information ‘hiding’ in these files if using normal search tools, stresses again the need for proper metadata tagging.

Conclusions

We can conclude that an inescapable condition of a consistent, transparent knowledge management at the ECDC is to establish a common, shared metadata structure for all information objects.

We suggest that an appropriate metadata structure can be created using the Dublin Core Metadata Initiative [3]. Besides the ‘mundane’ metadata as contributor etc., the content–specific metadata elements[1] are the most crucial.

ECDC has to run in the integrated EU database of surveillance networks of communicable diseases, a Threat Tracking Tool that monitors suspicious events, to publish Eurosurveillance etc.

Variables of all the mentioned databases and ECDC texts can nd will provide the core set of terms that is under evaluation of the ECDC expert staff.

[1] in DCMI terms: subjects and keywords

Conclusions – continued ....

The folder structure of the document library can be seen also as metadata characterizing the information object, a typical example is ‘Directory of R:\Epidemic Intelligence \ RiskAssessment \ Guidelines\Retrospective on US Health Risk Assessment_files’.

These strings are actually describing the documents with characteristics that are not necessarily availably from within the object. Therefore folder / subfolder names are “subject metadata”.

All these resulting terms will be collected and will be used to establish the ECDC core terminology, behind that a ‘core ontology’ will serve to maintain consistency and transparency of the terms set as well.

Summing up our future plans, we want to progress towards:

- establishing **a core ECDC terminology, with ontological basis**, later to be extended to a full terminology
- using core terms as **a metadata set applied to all information entities**, remaining committed to standard document formats;
- Using a core ontology as a source for **a unified search function,** making the entire ECDC knowledge space transparent, easy to navigate and available to the wide variety of end users.

References

[1] Miller GA. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, The Psychological Review, 1956: 63, pp. 81-97

[2] Handschuh S, Staab S, Maedche A. CREAM: creating relational metadata with a component-based, ontology-driven annotation framework . In: Proceedings of the 1st international conference on Knowledge capture, ACM Press, Victoria, British Columbia, Canada, 2001, pp 76-83

[3] DCMI Usage Board DCMI metadata terms. http://dublincore.org/documents/2003/03/04/dcmi-terms/ , 2003.

Address for correspondence:

László Balkányi MD. PhD.

E-mail: laszlo.balkanyi@ecdc.europa.eu

European Centre for Disease Prevention and Control (ECDC)
Tomtebodavägen 11A, SE-171 83 Stockholm, Sweden