MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

A Case Report of First Steps Taken at an European Agency, ECDC to Introduce Knowledge Management Services (description of knowledge sources and 'extraction' of metadata in practice)

László Balkányi[a]

[a] *Unit of Scientific Advice, European Centre for Disease Prevention and Control , Stockholm, Sweden*

Abstract

As part of efforts toward building knowledge based services at the new European Centre for Disease Prevention and Control (ECDC), authors investigated content and extracted metadata from a large document base.

ECDC 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 provisional, start-up state, it was found that a set of systematic measures among them notably introduction of metadata enhanced templates (using Dublin Core NSMetadata 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.

Keywords:

information management, public health informatics, communicable disease control, knowledge based services, document management, standard tools, European Centre For Disease Prevention and Control, ECDC

Introduction

This paper describes the first steps taken at a European agency, European Centre for Disease Prevention and Control, ECDC to introduce knowledge management (KM) services. Describing the chain of thoughts while planning the services and the experiences of the analytic work hopefully help others with similar tasks to understand the nature of information handling at their organization and to plan better KM services.

ECDC is an EU agency that has been created to help strengthen Europe's defense against infectious diseases. The mission of ECDC is to identify, assess and communicate current and emerging threats to human health posed by infectious diseases. To achieve this mission, ECDC works in partnership with national health protection bodies across Europe. ECDC aims to pool Europe's health knowledge, so as to develop authoritative scientific opinions about the risks posed by current and emerging infectious diseases. The Centre, currently staffed with ~ 80 experts and management, will have an over € 50 million budget by 2010 and its staff will grow to 300 over the coming years.

Establishing a knowledge-base (KB) and a set of knowledge-based services (KBS) is clearly a key issue for the full blown operation of ECDC. KBS's developed for human- and KBS's developed for machine (that is software) users will serve all activities / all units of ECDC.

ECDC has a matrix organization with three technical units, an administrative unit and seven horizontal disease-specific projects. The three technical units are the following: The Scientific Advice Unit (SAU), the Unit for Preparedness and Response (PRU), and the Surveillance and Communication Unit (SCU). SAU's main tasks are to provide sound and independent technical and scientific advice; do and maintain an inventory of communicable disease related resources; build an operational network of experts and scientists in Europe; build further on the scientific competence and experience of the scientific community for communicable diseases, actively participate in all key scientific conferences and meetings. PRU's main task is to keep track of emerging health threats inside and outside the EU, provide rapid risk assessment and coordinate a timely response to such threats, done by operating the Early Warning and Response System (EWRS), organizing training programs for public health, and supporting outbreak missions of MS experts. SCU's main task is to take over and operate the surveillance of communicable diseases at the EU level from the designated surveillance networks. Data will be analyzed and the information disseminated to those who need it for public health decision taking in Europe. The ECDC disease specific activities are organized within seven horizontal projects with team members from all technical units: 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 nosocomial infections.

In ECDC we plan to establish a knowledge base for human users to serve as electronic information source, enhanced / supported by specific services i.e. knowledge mapping, intelligent search, cross-referencing, and other tools to navigate in the information space. It might be equally, or on the long run even more important to establish KB services that can be used also directly by other information systems, the surveillance networks, the Early Warning and Response System such as terminology services and later also by other advanced systems / functions, enabling high level, standardized reporting, (semi-) automated extraction of critical information, semantic crosschecking of information gained form independent sources, etc.

At planning these services, we want to avoid 'reinventing the wheel' and as a strategic decision, we want to stick to international standards ensuring interoperability and long run transparency of emerging systems/modules/functions. Fortunately, recently, a number of general and medical informatics knowledge representation and handling tools became more and more mature, and available as tools for such tasks on several levels from data formats to knowledge representation and process & system modeling (the XML family, UML, OWL, etc). The buzzwords 'semantic web' and 'web 2.0' describe a set of emerging tools / methods that should be considered at planning a future proof solution portfolio.

The below described work is one of the first steps taken toward the realization of KM services at ECDC. A knowledge needs assessment revealed that there is a pressing need for better information retrieval, handling of the already existing and rapidly growing electronic document set of ECDC. It is clear, that on the long run, as part of the integrated ICT services, ECDC will have to have an integrated electronic content management system dealing with the internal documents and also the ever-growing, mostly open partly restricted web content. Planning, procuring and implementing such a system takes at least 1,5 – 2 years, so we needed an interim solution, that serves also as a sort of sandbox for the final application. We describe here the analytic parts of the work, the thinking about metadata, some innovative methods we choose and show the results of these preparation tasks.

Materials and methods

Materials

Our objects, materials to study were the electronic documents accumulated on the drives of ECDC since its operations started in May 2005. The 'snapshot' of the files – folders structure was taken in August, 2006, after roughly one and half year of ECDC operations. Only files and folders of the technical (scientific) units were analyzed – that of the administration, financing, human resources and routine emailing were not drawn into the analysis. All files generated during this period were taken in account, classified into documents[1], spreadsheets, presentations, and 'all others' categories.

Methods

To prepare later seamless introduction of content management, we decided to take three steps

- an analytical step to understand the build-up process, the accumulation of information
- extraction of metadata and setting up a keyword library
- establishing a set of metadata "augmented" documents

This paper focuses on the first and on the second, closely related steps of this process.

To achieve the goals, simple tools were used as follows:

- command line tools and disk cataloging utilities for extraction of file / folder structures
- a Unicode text editor for building/editing xml files
- an online word processor for producing ODF compliant document templates
- a spreadsheet application for file statistics

Results

ECDC stores its electronic documents on file servers, where a shared space (a 'drive') was set up to store documents that should be reached across all the units, and 'unit specific' spaces (virtual 'unit drives') were assigned to the three above mentioned units, the SAU, the PRU and the SCU. The data characterizing file accumulations and distribution of the first one and a half year were analyzed in depth. The following numbers just summarize the most characteristic aspects. The document distribution in folders was investigated with the following results – see table1:

Table 1 - The number of folders/ subfolders existing at each level of the folder- tree

Subfolder depth level:	No. of folders in "shared" drive	PRU drive	SAU drive	SCU 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

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

Subfolder depth level:	No. of folders in "shared" drive	PRU drive	SAU drive	SCU drive
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 no. of folders:	**4615**			

These numbers and some others following below are analyzed in the discussion section of this paper. It has to be noted, that the folder structure is the result of a non-regulated, 'organic' growth of the document set in a just formed, new organization, where highly autonomous and highly educated experts worked mostly either alone or in ever changing, problem oriented, ad-hoc teams on emerging problems. Roles and tasks were not yet 'frozen', most experts did extensive multi-tasking and took part also in building up activities. Information communication technology services were focused on providing a stable, robust office services background and electronic communication services, with no emphasis on electronic workflow support or organization of content – due to lack of capacities. However the amount of accumulating documents was / is staggering – be aware that the value axis below is logarithmic! See figure 1.

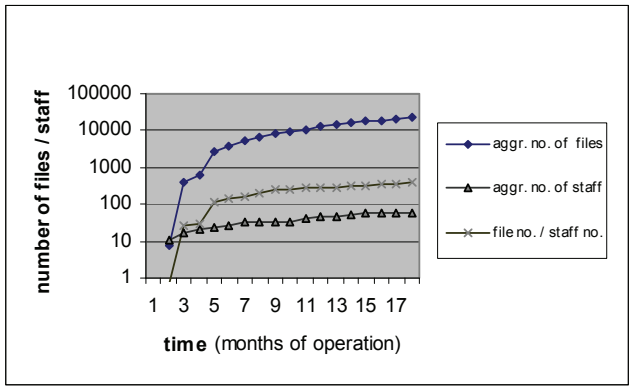

Figure 1 - The growth of staff and the aggregated numbers of scientific document files on ECDC servers

The distribution of files is detailed according to the most frequently used file types in table 2.

Table 2 - file type distribution

location of files		documents	spreadsheets	docs to publish (pdf-s)	presentations	miscellaneous (jpgs, html, txt, rtf, etc.)	all
shared 'drive'	abs	5797	557	4048	1167	6695	18264
	rel	32%	3%	22%	6%	37%	100%
PRU 'drive'	abs	2508	328	897	390	3337	7460
	rel	34%	4%	12%	5%	45%	100%
SAU 'drive'	abs	880	758	163	58	1618	3477
	rel	25%	22%	5%	2%	47%	100%
SCU 'drive'	abs	2022	609	341	149	1926	5047
	rel	40%	12%	7%	3%	38%	100%
total	**abs**	**9185**	**1643**	**5108**	**1615**	**11650**	**29201**
	Rel	**31%**	**6%**	**17%**	**6%**	**40%**	**100%**

Discussion

Discussion of the results focuses on understanding and interpreting what is behind the numbers.

First, the files / folders structure: It is obvious, that it is almost hopeless to navigate efficiently (i.e. quickly and easily) in such a tree. The depths of more than ten levels and the (sub-)folder numbers ranging from one to more than six hundred (on depth level 5 of the 'shared drive) is remarkable. According to a classic, widely accepted paper [1] of cognitive science the pan of short time memory allows us to handle 7(+-2) "chunks" of information at any give time, due to inherent limitations of short term memory. Working in complex folder hierarchies, we use short term memory, therefore this well documented phenomenon should push us to reorganize our file/folder structure into embedded hierarchies of roughly 7 by 7 matrices (not much more than 7(+- 2) subfolders in a folder and not much more than 7(+-2) files in folder). The material, we deal with, however does not allow such a grouping. Concerning searching and retrieval in such a large, complex information space, does a simple or even an enhanced "brute force" search engine help? We think not, as desktop search engines have a significant limitation: they can search only by the string existing within the information object. If we have thousands of documents on very similar topics, the search by strings (words or parts of them) will be never specific enough to find a particular information object among them. Such a search will also be not helpful to find a set of related documents, connected not just by having the same words, but by some other characteristics, as e.g. referring to a geographic region, whilst in the text only countries or cities are named (e.g. in case of an out-

break investigation). Using metadata is of course the answer to this problem [2]. If a well maintained keyword vocabulary is used to tag the content and other characteristics of the documents, such problems can be solved. Even better if the 'vocabulary' is based on a logically sound ontology, meaning that later on we will have the option of semantic operations, of complex searches at hand. Before discussing the metadata issue, it is worth to discuss *a second aspect: the growth of the information heap*, resulting in almost thirty thousand information objects in one and half years, as seen on figure 2. below.

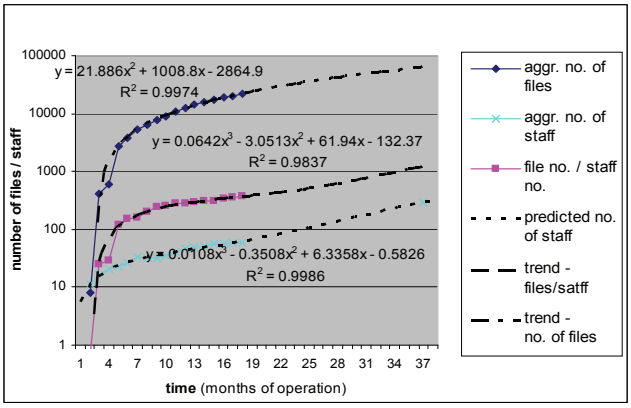

Figure 2 - trends of the growth of staff and the aggregated numbers of scientific document files on ECDC servers

ECDC will grow to the predicted staff of ~ 300 in the coming two years. Using this number trends can be examined concerning the future number of files and the file to be handled by a given staff member. Well fitting trend lines show that the number of files will reach nearly a hundred thousand and each staff member will be 'responsible' for more than a thousand file. This predictions might be disputed as more and more 'services' of ECDC will become 'operational': like the surveillance networks and the organization will turn to a 'normal operation' mode from the current build–up period. However most of the functions 'producing' information are already functional, even if only in skeletal format, and the curves show that the characteristics of the buildup period (sort of 'combinatory growth' of staff & information) became already more 'leveled' after the first nine months.

The last, third issue before we can discuss the required metadata in details, is the *distribution of files types* and its consequences: *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 other, special files used to develop databases and applications. 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.

As a result of the discussed facts, 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. In keeping with our commitment to internationally accepted standards and protocols, we think, that an appropriate metadata structure can be created using the Dublin Core Metadata Initiative [3]. The goal of this initiative is to standardize the names of the metadata elements with which an object is tagged. This results in great benefits with respect to electronic processing, since the metadata set of an object will be portable across different systems. As the DCMI matures, more DCMI-aware applications will become available, increasing the usefulness of the metadata.

As a preparation step to managing metadata, a set of document templates has been provided which incorporate the recommended metadata structure. The templates were selected to cover a set of common documents which are created at the ECDC on a regular basis. The specific types were selected based on the observed recurrence of documents in the shared folders and using feedback from semiformal interviews. The motivation for using these templates is that a significant subset of the metadata will be preset for each document. Since the templates are linked to each other based on a principle of inheritance, creating new templates is a more straightforward process. In other words, an SCU mission report template was created by extending the generic mission report template which was in turn created by extending the basic ECDC root template. As a result, all templates inherit the metadata of their parent template and add new elements specific to their additional characteristics. By using templates staff members will be relieved from the bulk of metadata tagging, only content specific metadata (akin to traditional 'keywording') has to be added.

What are the sources of these content–specific metadata elements, in DCMI terms: subjects and keywords?

Term Name: subject	
URI:	http://purl.org/dc/elements/1.1/subject
Label:	Subject and Keywords
Definition:	The topic of the content of the resource.

Figure 3 - DCMI definition for content specific metadata

ECDC has to run in the future among others the integrated database of surveillance networks of communicable diseases operated for the EU, the Threat Tracking Tool that monitors suspicious events, publishes Eurosurveillance, a periodical that has many decades experience in keywording papers related to communicable diseases. Variables of all the mentioned databases and the corpus of Eurosurveillance provided the core set of key word list that is under

evaluation of the ECDC expert staff at the time of writing this paper.

However, the folder/subfolder/filename structure of the present document library can be seen also as metadata characterizing the information object: See some typical examples:

> 'Directory of R:\Epidemic Intelligence \ RiskAssessment \ Guidelines\Retrospective on U_S_ Health Risk Assessment_files'

or

> 'Directory of T:\Surveillance Networks\+Evaluation docs\Evaluation documents\Web Questionnaires\Epi-ContactPoint.aspx_files'

These character strings, although presently part of an arbitrary folder structure are actually describing the documents with features, characteristics that are not necessarily availably from within the object. Therefore we are using the folder (subfolder names as candidates for future keywords and also they will be used as "subject metadata" in a later planned migration to a metadata using content management system. Besides parsing there is obviously a data-cleaning, filtering step, that has to be done, finding redundancies, correcting misspellings etc. and establishing preferred set of terms.

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

Conclusions

Numerous off the shelf solutions exist which could address the issues of knowledge management in ECDC. Software such as document management systems and virtual file systems will at least partially solve these tasks. However, they are designed mostly to integrate into a business workflow model and address archival requirements which differ from an organization such as the ECDC. Largely, the solutions available are still document-centric and target a different set of priorities. As the official European source for information on, and related to, communicable diseases, the ECDC has a specific set of requirements which differ from those of both private companies and academic institutions. ECDC needs

- efficient data archival and retrieval:

Unlike information such as accounting data, processed and stored using well-defined methods, the scientific information in ECDC is more open-ended and has more complex inter-relationships, requiring something more feature-rich than a basic filing system. The experts working at the ECDC will not be primarily concerned with the administrative processes of archiving and retrieving information, but with the content. Archiving and retrieval should happen as transparently as possible, reducing procedural overhead. Unlike academic institutions, ECDC may not always have a single standard mode of operation. During, for example, a serious outbreak, information may need to be found, processed and stored within much narrower time constraints ("peace mode" – "war mode" procedures)

- easy referencing of information:

A document-centric approach solves the problem of relating a set of data with a set of conclusions by embedding both into a single document. Inter-document references are handled using citations the most efficient of which make use of identifiers such as ISBN or ISSN numbers. For ECDC an equally unambiguous solution is required but intended for a more heterogeneous knowledge management solution, which goes even further, allowing many-to-many relationships. The method of referencing information must also be usable by both human and automated agents (this follows from the previous requirement).

- a long information life cycle:

While the length of information retention at an average IT company may be measured in years, the length of information retention at the ECDC may be estimated in decades (in this case we are not considering backups, but information stored in primary storage and readily accessible). This includes the raw and processed data, the ability to flexibly process the data, and any references that exist between sets of data and between data and analysis results and conclusions. This profoundly affects all of the information requirements. We must consider that most of the popular file formats in use today are only a few years old and that users are already experiencing problems when attempting to read files only two decades old. In order to address these requirements in a way that will provide the most benefit to the organization in the years and decades to come, some groundwork needs to be laid out.

Policies and mechanisms are required which standardize the following across all units and horizontal activities:

- establishing and using a core ECDC terminology, with ontological basis
- establishing a metadata set which can be applied in whole or in part to all individual information entities and which will incorporate also the keyword set
- a well-chosen set of acceptable data formats for use in the knowledge repository

  (Note: there is a need to decouple of data formats from the applications used to read and write the data. Any file which can only be read using a particular proprietary application is not suitable for an information life cycle of the length which we are considering here)

- having a commitment to formats and protocols conforming to internationally accepted standards

Summing up our experiences, an envisioned implementation for managing the information contained by the internal electronic document set will incorporate several key methods and tools:

- using open document format and general XML encoding
- using metadata embedded into documents for redundancy (using metadata enriched templates)
- establishing a database which will hold the extracted metadata for processing during normal execution
- using controlled vocabularies with ontology 'back-up' to enforce consistency
- establishing a unified search interface making the entire knowledge space available to the end users.

We hope, that the above presented experiences and ideas will help others to establish similar services at different organizations.

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.
European Centre for Disease Prevention and Control (ECDC)
Tomtebodavägen 11A, SE-171 83 Stockholm, Sweden
E-mail:Laszlo.Balkanyi@ecdc.eu.int