UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG
formation continue

CERTIFICAT EN GESTION DE DOCUMENTATION ET DE BIBLIOTHÈQUE
2006-2007

Specifications for a document repository tool for a drug safety department

Have data? Then need data management!

Sandra Manzi Kutner
Genève

(Septembre 2007)

TRAVAIL FINAL DE CERTIFICAT

Déposé auprès de

Michèle Furer, Service de coordination des bibliothèques, Université de Genève et
Corrado Pettenati, Directorate Services Unit, Cern, Genève,
responsables scientifiques du module “Gestion des nouvelles technologies de l'information”

Jean-Blaise Claivaz, Service de coordination des bibliothèques, Université de Genève,
personne de référence pour le suivi pédagogique du travail

Table of contents

Abstract

Literature tracking is essential to research groups that need to keep informed and up-to-date of the team's readings and findings. Specific software can address this need, and a variety of solutions are available on the market.

In this dissertation, we describe how some of these solutions can be adopted by a pharmaceutical company's department, and what considerations we need to formulate and analyze in order to satisfy user's and company's needs.

Introduction

Researchers who consult scientific documentation need to track their readings, in an orderly manner. This is particularly true for a pharmaceutical company's department dealing with drug safety, where a few dozen specialists need to track published literature in relation with their products' therapeutic area, not only on a personal level, but also to share information with colleagues.

MerckSerono Global Drug Safety (GDS) department is in this respect typical, and we shall describe the current system for tracking literature, and highlight the advantages of migrating from the current listing in favour of a more appropriate solution.

In view of this migration, we have contacted key users in order to establish their needs and discuss possible improvements. Major desired features emerged, such as a customization, ease of consultation and searching, citation facility, and especially definition of controlled keywords.

At the time of writing, a project team is evaluating options for a company-wide solution, where GDS is part of the definition team. This dissertation benefited from discussions with users and other information resource specialists, but the responsibility of anything stated lies entirely on the author.

Note: throughout this paper, we refer to "users" in opposition to "administrators" of a proposed tool. Users are staff members that can view the data, while administrators are the ones that can add or modify data in the tool.

Drug Safety and Bibliographic Data

As part of the Development function of a pharmaceutical company, MerckSerono's Drug Safety department ensures that drugs are monitored with respect to any safety issues, as well as analysing possible emerging trends and interacting with authorities to report relevant adverse drug reactions. The specialists are usually medical doctors, pharmacists or have training in other life sciences.

Adverse events are either collected from clinical trials, or spontaneously notified from the post marketing setting, or are actively searched by GDS in the published literature. Generally speaking, the GDS scientists deal with adverse events at two levels:

1. analysis, follow-up and reporting of individual events that occur with drugs either pre- or post- marketing. This analysis considers medical plausibility based on the mode of action and safety profile of products in the same class;

2. analysis of clusters of events taking into account the population perspective (epidemiological data).

In this setting, referring to findings by other researches is a necessity that no-one can overlook. In its turn, this gives rise to the necessity of collecting and tracking the consulted literature in a structured way. Articles that reported adverse events need to be kept for future reference, and are to be shown to health authorities as part of drug follow-up.

Besides ad-hoc responses to events, at regular intervals pharmaceutical companies are requested by health authorities to critically review drugs already on the market, and deliver a periodic safety update report (PSUR). PSURs list extensively literature search with respect to the given drug, including those same articles that were used for detecting adverse events, as well as other scientific publications related to the drug and to the indication supporting the discussion in the PSUR.

It is therefore useful and highly appreciated to be able to cite those articles without being hindered by bibliography formatting and word processor cut-and-paste, considering that for these reports articles cited are between 20 and 60.

As nowadays the near-totality of scientific literature is available in electronic form, the department's library consists mainly of an electronic collection of PDF files, extracted from scientific journals.

In a sense, this article collection makes up for the department's digital library, where the acquisition policy is dictated by the specialists' need to follow up on the company's drugs and therapeutic areas. Obviously, having such a collection is useful only if there is a means to access in a timely and appropriate manner this information, a task that is normally performed by a (library) catalogue.

As it is often the case in library information services, data regarding a printed object is entered by one authority, and then generally imported by others in their own catalogues or database management systems.

In this field, PubMed, a service from the US Library of Medicine (National Library of Medicine, NLM), gives access via a web interface to the majority of published material in the field. Journals, conferences and other published material are also usually exportable to a citation tool, or at least to a common index by the NLM. Similarly, all journal

publishers give the possibility to export article citations to a file, very often giving the choice of a few formats or bibliographic management tools.

Having a common base for bibliographic references not only makes it easier to add new entries to a catalogue, it also limits errors (e.g., typing, spelling, omissions, etc.) that would be frequent in such subject area. In addition, since PubMed assigns an identification number to each article, this can be exploited for rapidity of access and cross referencing.

What is in place now?

Currently, a listing in tabular form exists, that has been in place since 2002. This solution is a plain flat listing, which fulfills the need of identifying which PDF file corresponds to which bibliographical entry. A department-specific article coding was put in place, which classifies articles according to the drug referred to or possible event, and if it was used as a reference in a report.

This listing does not, however, allow for data to be automatically imported via PubMed or any other provider, nor exported and cited as bibliographic references in one's report.

When a specialist requests an article, the documentalist enters the meta-data supplied by the requester, proceeds to obtain the article and once the article is obtained, a link to the local PDF file is also added. [1]

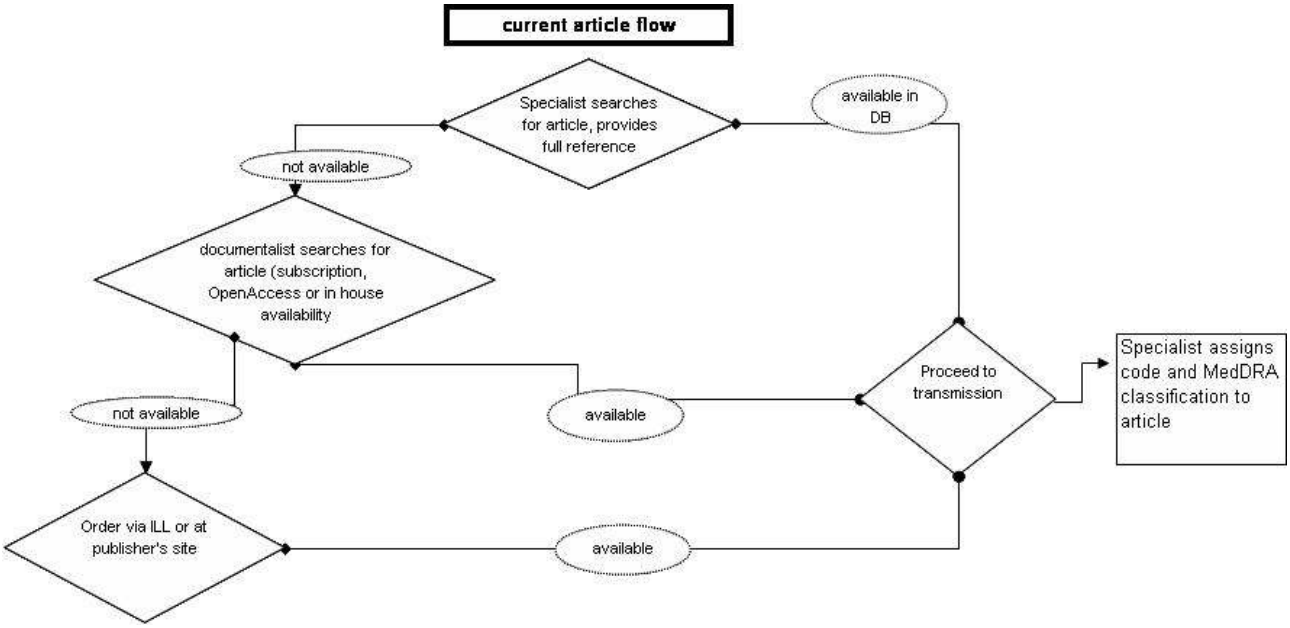

Figure 1: Current article search and distribution flow

To sum up, the flow in place is the following:

[1] The cases of articles not being obtained are extremely rare, and we do not therefore analyze this possibility any further.

- the documentalist enters bibliographical data (title, author, journal, issue, etc.) and the link to the PDF file;
- the specialist assigns the code, hands it on to the documentalist for the classification of the corresponding article;
- any specialist can access this listing, search through it (one term at a time only), view the references contained in it and any PDF file it links to.

The classification in place assigns terms according to a department-specific nomenclature. Below is a snapshot of these fields and their significance:

Class	Sub-Class	Source	Report	Key Words
which product, substance, or class of illness is being reported here (e.g., “23” for Interferon-beta, Rebif)	MedDRA[2] classification of adverse events (a 2-digit defining a system-organ class, e.g., “21” for reproductive health)	if the requester wants to keep information about how the article was obtained, (e.g., via local representative, etc.)	if the article is linked to a PSUR or to a DER	any other keyword not present in the title or implicit in class / sub-class (non-controlled terms)

Table 1 Article listing's current classification

The class and sub-class coding have a double duty. On the one hand they are assigned by the specialists, and can be used by them to retrieve or identify one or more articles. On the other, when the system was put into place in 2002, only about half of the articles were available as PDF files; there was therefore a need to file paper copies, and this classification was suitable for the task.

Since its introduction, this listing initially served the department’s needs well. More recently, though, considering the increase of articles, the listing approach shows its limits and shortcomings. Users have trouble sieving through a few thousands entries, the software does not enable multiple term searches (e.g., author=smith and title=cardiac and year=2004), which makes searching lengthy and tiresome.

Moreover, citation features are critically needed for writing (sometimes) lengthy bibliographies, and the availability of citation software makes the current approach redundant.

Good search functionalities are also a major request, since users are very literate in database search and can manage sometimes complex queries.

In January 2007, Serono merged with Merck Ethicals Division, which brought two separate Drug Safety departments to integrate and harmonize their processes.

This subsequent larger need for documentation tracking further motivated the necessity for a change. Merck’s Safety in Darmstadt was also in search of an article tracking, or document repository, but had not constructed a listing of previous articles

[2] Please refer to the acronyms listing at the end of the paper.

within the department.

Serono's listing currently holds ca. 4300 articles, starting from 2002. In the period January to June 2007, circa 400 entries have been added.

The status quo is not satisfactory for the reasons listed above: bad search facilities, lengthy data entry, error insertion, possible work duplication, non-existent citation facilities..

Changing for the better

How are we going to go about it? What is our wish list? What do we need to foreseeably include in this tool? What do the users want?

First of all, we enquired about the original thought behind the solution put in place in 2002, we gathered input from a focus group, surveyed the other in-house software solutions and solicited advice from in-house information specialists.

We came to consider migrating the listing to a bibliography management tool, and during our discussions with other departments we came to learn of a more company-wide solution being put into place, which we describe in Section Options investigated.

A company-wide common repository would allow for a number of specialists in different departments to access the same articles without creating duplication of effort, data, and files. In addition, given the interaction that some departments have, this would improve flow of information.

Focus Group Survey

The specialists who have been using the listing for at least 3 years find the status quo is useful, but would certainly welcome integrated solutions with citation importation for reports and better search functionalities.

As reported in Table 1, specialists rely on the classification entries, and when these are left empty they rely on author's name or title of the article. As mentioned,

- the coding is important only if paper-only copy exists of article (in which case the coding is the cataloguing reference for the article, being kept in a hanging file folder)
- the coding is important for DER tracking
- the coding is important for PSUR or any other report tracking

The survey with the focus group brought the conclusion that a literature database was indeed much needed.

As the current system contains a number of fields dedicated to the department's classification, a similar flexibility would be sought in the new tool. These specific fields are especially important to link adverse cases to articles, and to re-access previously used literature for further use (e.g. for PSUR).

It remains to be defined what type of tool would be selected for the task, and the number of articles that would be migrated to the new tool.

Users' information literacy

The sought tool is going to be used by users that have a fairly good level of information literacy, who know when to rely on standard searches and when to resort to advanced search, and navigate between standard search terms (author, title, journal) and company-specific terms.

Possibly, users will be accessing the database via a querying system on a web interface, while the database is to be fed by information specialists or external vendors.

Options investigated

Essentially, four types of tools can be taken into consideration.

1. custom-made database
the choice is vast, it involves a lot of work and implementation,

2. off-the-shelf bibliographic database or bibliography management tool (BMT)
here the choice is less vast, and may include proprietary and open-source tools

3. document management systems (DMS)
proprietary and open source tools exist; a big drawback is in the citation management and import of entries.

4. product literature database (PLD)
internal or fully or semi-outsourced solution, whereby a vendor manages the database, both feeding it with new articles and sending alerts to the specialist; it is a specialized database that pre-digests newly published material regarding one specific drug or substance.

Within the company, a project is underway to select a tool, and GDS is collaborating with other departments with similar literature needs on specifications and choices.

It is necessary to keep in mind that an off-the-shelf bibliographical management tool is often conceived for a single user, or for a limited number of users, or for a limited number of articles. The needs of dozens of users (if not in the hundreds), who are very large consumers of scientific literature, might not be met by a majority of these tools[3].

In schematic form, we have three objects: the PDF file, the PubMed (or any other publicly available) meta-data describing it, and the department-specific information. The clear importance and advantage of one system over another is in the flexibility it might offer to include the department-specific data, since this is a feature that really makes the tool apt for the job, and, more importantly, one of the requests from the focus group.

[3] As an example, some citation tools have a limit to 50,000 items, which is already the current holding of the Medical Information department. For a company-wide solution, this figure would be attained in less than one year.

Department-specific data is a controlled vocabulary set or code (the drug class, the MedDRA system organ class), but also semi-controlled data (for instance, the type of report the reference was used for, or some descriptive text regarding the article).

A custom-made database solution would require some heavy work from programmers within the company or from consultants. It would certainly allow for high customization, but would also bring debugging and system maintenance with it, both rather time- and energy-consuming tasks. Even if this option is not discarded right away, the downsides involved in setting up such a piece of software, considering the availability of commercially viable options, do not make this solution a highly likely candidate.

Likewise, a document management system, originally intended for sharing documents across different users and sites, is more geared towards tracking revisions of documents and version control. It can be used as a simple repository of articles, and would allow a lot of flexibility for descriptive fields. Two major drawbacks of a DMS are the lack of the citation facility, and the unlikely importing of bibliographical data from PubMed or any other provider, again central to user's requests.

By elimination, two final classes of products are therefore in line: a bibliography management tool and a product literature database.

What is a bibliography management tool?

A BMT is a piece of software that allows writers to keep track of their literature search and work, making it easier to include specific citations in a word processor, without dealing too much with formatting; moreover, a BMT often acts as a database for all articles used, allowing for searching and retrieving not only the meta-data but usually also a locally-accessible PDF file of the article in question. From this point of view, a BMT acts as a document repository.[4]

One of BMTs main features is the possibility to import citations (e.g., from an on-line catalogue or other databases), which makes feeding the tool itself less cumbersome and avoids data-entry errors. Also noteworthy the possibility to include personal notes, annotations on the record itself, that is later visible as part of the article's entry.

Specific questions arise with respect of this department's needs. These include:

- the need for customizable fields (see above for MedDRA classes)
- accessibility on a number of sites (Geneva, Darmstadt, and possibly other corporate locations).

For our specific purposes, it would be wise to customize the user's annotation facility within specific boundaries.

Indeed, in most off-the-shelf applications this field is a free-text area, where usually the user jots down his own remarks on a document, like one would do in pencil on the margin of a paper document. It is a useful feature, especially for the solo researcher who enters data in the tool, manages and uses it by himself. However, as soon as the tool is to be used by a number of different people (even a potentially large number, say one hundred or so), the need for a more structured and controlled set of data for this field

[4] Of the number of tools available, the bibliography lists some reviews (open access articles) that highlight each tool's functionalities and specificity. Since we did not however carry out an on-site test ourselves, we refer the reader to these articles or to the product's web site for information.

becomes evident.

If we want to allow free-text annotation, we should investigate for more than one annotation field, so that controlled data can be entered in one (aptly named something like "MedDRA information", for instance, or "GDS classification"), and more informal descriptors (e.g., "cited by Brown and Smith, 2005") should be entered in a separate field.

What is a product literature database (PLD)?

These tools have entered the market lately, and go loosely under the name of "product literature database". They are specifically geared to the pharmaceutical companies' needs to constitute a database listing all scientific literature published in relation to a specific product or substance. An outside vendor is generally entrusted to scan regularly for newly published material regarding a specific drug, classify the findings in a pre-established manner (e.g., by drug interaction, by dosage, by geographical distribution, and so on), and supply both the article itself and the full classification meta-data.

PLDs being rather new products, they are not very well known, few vendors are on the market for the time being, but they can offer the flexibility that usually comes with novelty.

The advantage of this type of solution is that it can provide full-text search in all the articles it has collected, so that highly narrowed queries can be formulated.

A second advantage is the availability in the database of articles published by any journal, regardless of its publisher or its open access availability[5].

Starting from the assumption that a specialist is interested primarily in one drug and its interactions, a PLD groups up articles by substance or drug name (the "product"). In this way, searching for example for "cardiovascular diseases" will scope only on this class of diseases and the chosen drug.

This operation provides not only the economy of typing a search term. In scientific articles, a drug name can be present in the title, in the abstract or in the full text; under its marked brand name or as a substance; sometimes it is assimilated to a class of drugs (e.g., corticoids). PLD vendors carry out a mapping of a number of terms, strongly in relation with one specific drug, and gather under one single term articles in relation with the company's product.

PLDs can also implement full-text search on articles available in the database. This can be viewed as both a strength and a complication. Full text search can retrieve otherwise unfound sources, since certain terms would not show up as keywords or in the title/abstract part of the article. On the other hand, this strategy is going to lead to the retrieval of articles that are potentially of little relevance to the original query (creating "noise").

Therefore, it might be advisable to include full-text search as an optional functionality, for instance as a section button in the advance search interface.

[5] Copyright issues are not the focus of this work, and we need content ourselves with the assurance that the vendors are to acquire copyright permission from holders.

Considerations common to all options

Whatever the new tool will be, a few considerations should be common. These are intended to make stakeholders more at ease with the tool at first encounter, and thus favouring use of the tool and assuring its actual usefulness.

Nobody disputes the advantages of having a web interface, since the browser's functionalities are acquired for free, and users are well-accustomed with this tool already. In addition, since databases work typically on a client/server architecture and this also is a typical web interface, putting them together comes rather naturally.

A second consideration is the possibility to export the database to a different formatting, in order to secure the repository from software obsolescence, or any other reason that might make accessing it no longer possible. This links with the web-interface aspect, which should facilitate exporting.

Access to the database on various corporate sites, prompt update of the database's indexes, especially when an outside vendor is involved, are also worthwhile considerations.

For the search functionality, a "basic" Boolean search should be offered, together with a more advanced one. In this latter case, nested searches might be useful (e.g., entering as search criteria ((lung AND infection) OR pneumonia). This search type would be especially useful if full-text access on the PDF files is made possible, but not so vital if all search criteria can only rely on the article's meta-data.

Apart from these, we take for granted that common search functionalities by title, author's name, date, journal, etc. will be available.

As mentioned earlier, the ease with which references can be cited is crucial. Exporting bibliographical references in a pre-defined format with a few mouse clicks is a task at which the tool should excel, enabling users to include references in a word processor in different referencing formatting styles.

It is however unclear if the tool will benefit from compliance with the Z39.50 protocol for cataloguing. This protocol intends to make possible the communication between databases that are linked via an internet connection. It is much more geared to integrated library systems than to bibliographic or citation tools. However, since part of the tool is to deal with importing citations and this protocol deals specifically with data exchange, conformity with this protocol might be a precaution to keep in mind.

Looking back to Serono's original listing, and regardless of the tool that will be selected, the following fields are to be considered:

standard features (title, authors, source, date)

medical literature specific identifiers: PubMed unique identifier (PMUI), other medical databases identification number; note however that not all articles have a PubMed id, e.g., not all non-English language articles are listed; if implemented, we have to allow for this field to remain empty.

For the articles that do have a PMUI, it is a convenience to be able to re-enter the number and, for instance, retrieve from that database related articles, or articles that cite the original paper, etc. One could arrive to this result in other ways (e.g., re-entering author and title), but accessing the information with one single code is a definite time saver.

MedDRA coding: its usefulness was crucial for paper-only articles, since it provided classification for storage; nowadays, since all articles are on-line PDF files, this aspect is not crucial for storage, but is still relevant for the specialist's searches.

electronic publishing identifiers: DOI (digital object identifier) codes are unique codes assigned by the publisher (or at least at the publisher's request). Because more and more articles are being extracted from the full publication's issue, the need arose to identify documents with a permanent code. In our case, it might be redundant to list both PMUI and DOI, and this double entry might just be a safety net for classification purposes. Note however that this field is absent from older articles (pre-1999) and possibly from articles received via interlibrary loan.

.

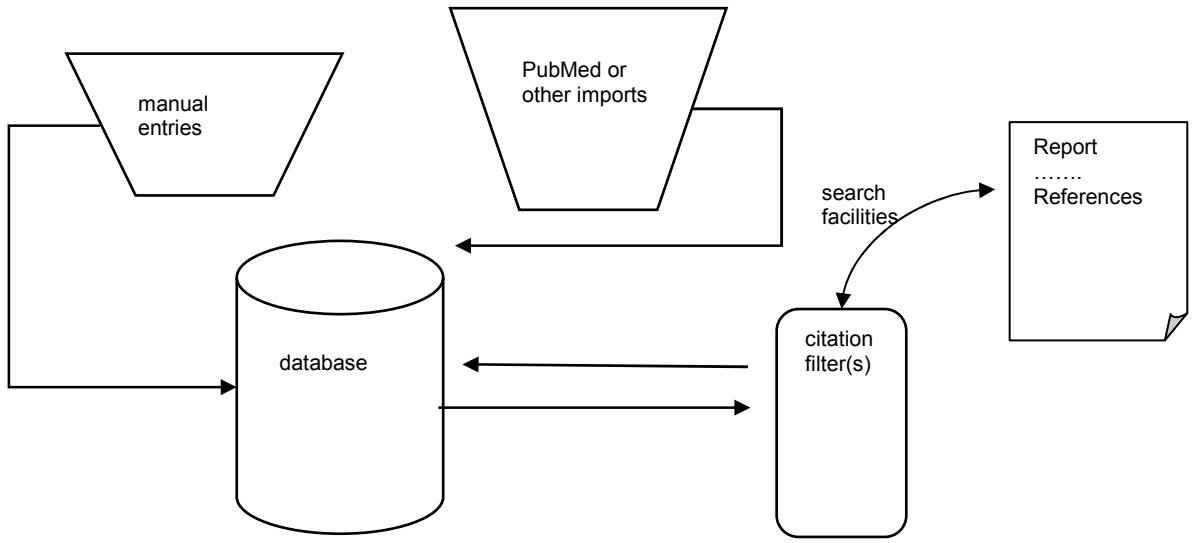

Figure 2: database citation structure

Access to this database would be open to all staff of the GDS department, identified via IP. Contrary to other pharmacovigilance data, where confidential information regarding patients can be present, data contained in published scientific articles is, by definition, public knowledge, and does not therefore require special protection.

Drug safety specialists are the primary consulting users of this database, but in practice any GDS staff member should be granted access to the data: for example, the data-entry clerk could need a reference for the authorities' report, a compliance specialist might need an article, as well as other non-regular users' requirements. This consideration is not a secondary one, in that it bears consequences on the user-friendliness of the design and interface.

The tool needs to be accessible to users in at least 3 "main" sites, i.e., Geneva (CH), Darmstadt (D) and Rockland (USA). However, ideally, it would be preferable that all users

worldwide who have a need to access this database could consult it at their desktop via the company's network. At the time of writing the feasibility of this aspect has not yet been ascertained, since a number of considerations beyond the scope of this dissertation need to be taken into account.

Conclusions

There is nothing very novel about setting up a database of articles, importing citations from a database, including references in a report, tracking articles previously read, and building a library of PDF files.

What we have outlined here is the complexity of choosing a tool that would allow different people in different sites to access such a database, and access not only the actual meta-data (i.e., the bibliographic reference), but also the PDF file it refers to, provide search criteria for in-house specific data (e.g., MedDRA coding, PSUR reference and the like), as well as classical fields (e.g., author, title, etc.). A company-specific, customized tool is to be selected and different types of users have been consulted regarding their requirements.

Although this report relates very much a work-in-progress state of affairs, it highlights the need for consulting the final users, survey the available options, analyzing advantages and disadvantages of each envisaged solution.

There is a conspicuous absence of reference to costs that any of the solutions would engender. One reason is that since the customization is such a significant part a PDL, its cost can vary depending on the final requirements. At this point of the project's life, we cannot therefore give any indications or use this aspect as a comparison vis-à-vis a BMT.

Finally, looking to the future and imagining migrating at least part if not the totality of the data, means that exporting the base (possibly via an independent computing language/platform) needs to be an investigated aspect or requirement.

Acknowledgements

My thanks go to Montserrat Miret and Alain Micaleff at Merck Serono's GDS, who allowed me to explore these needs within the department and to the many staff members for their input and feedback.

Jean-Blaise Claivaz, from the Geneva University, accepted to supervise my work; I am grateful for his insightful suggestions and feed-back.

Many thanks to Raoul Powlowski for proofreading the pre-final version, and helping me polish up the final text.

Last but not least, throughout this formative year, I have drawn strength from three generations of loved ones, who knowingly or unknowingly supported (and put up with) me. Endless thanks go to my parents, my husband and my children.

References

Standards

The International DOI Foundation, Washington & Oxford
DOI® System and Internet Identifier Specifications, Version 1.1
http://www.doi.org/factsheets/0607DOIIdentifierSpec1-1.pdf

The Z39.50 Document
http://www.loc.gov/z3950/agency/document.html

Product Reviews

Mani, Nandita S. EndNote X. J Med Libr Assoc. 2007 January; 95(1): 98–100.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1773034

Collins, Linda J. Reference Manager 10. J Med Libr Assoc. 2002 July; 90(3): 357–358.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=116418

Hendrix, Ingird C. RefWorks. J Med Libr Assoc. 2004 January; 92(1): 111–113.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=314118

Shapland, Maggie, Evaluation of Reference Management Software on NT. University of Bristol, 28 July 1999. Last modified: Apr 2001
http://eis.bris.ac.uk/~ccmjs/rmeval99.htm

Articles

Sable JH, Carlin BG, Andrews JE, Sievert MC. Creating local bibliographic databases: new tools for evidence-based health care. Bull Med Libr Assoc. 2000 Apr;88(2):139-44
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=35212

Lease Morgan, Eric et al., MyLibrary Manual. Vol. 1: Designing, Implementing, and Maintaining Digital Library Services and Collections with MyLibrary, by University Libraries of Notre Dame, 2006
http://dewey.library.nd.edu/mylibrary/manual/

Blog:

Comparison of Free Bibliographic Managers
http://mahbub.wordpress.com/2007/03/04/comparison-of-free-bibliographic-managers/

Product information:

Emscopes (Elsevier)
http://www.info.emscopes.com/brochure.pdf

Pi2 Solutions Ltd.
http://www.pi2solutions.com/backgrounder.asp

Kaim Informatics
http://www.kaim.com/altsite/menu_frame.pl?page=history/case_four.htm

EndNote
http://scientific.thomson.com/products/endnote/

Reference Manager
http://scientific.thomson.com/products/refman/

ProCite
http://scientific.thomson.com/products/procite/

RefWorks
http://www.refworks.com/content/products/content.asp

Acronyms

MS = MerckSerono

GDS = Global Drug Safety

CSP = Clinical Safety and pharmacovigilance

DER = drug event report

LDSO = Local Drug Safety Officer

PSUR = periodic safety update report

BMT = bibliography management tool

PubMed: “text-based search and retrieval system” (http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html)

PMUI = PubMed Unique Identifier

MedDRA (Medical Dictionary for Regulatory Activities)

(http://en.wikipedia.org/wiki/Meddra)