# CERL Thesaurus file

Claudia Fabian
Bayerische Staatsbibliothek

## 1. What is it?

The CERL Thesaurus file has been created since 1999. It is maintained by the Data Conversion Group in Göttingen. The format is UNIMARC "inspired". It consists of three parts: places, imprint names (printers and publishers) and personal names (authors, editors, translators). As it is still in the process of being built up, more extensions are under discussion, mainly for corporate authors and owners and provenance information.
The CERL Thesaurus exclusively serves - as the choice of these entities indicate - early printing, which in the context of the activities of the Consortium of European research libraries (CERL) means printing of the hand press book area, starting from the beginnings (1450) until round about 1830/40. The Thesaurus is thus closely related to the Consortium's Hand Press Book Database (HPB) containing bibliographic records for early printing coming from different European cataloguing projects in a single and "one stop shopping" database run by RLG. The construction of this database which contains today 1,5 million records is undertaken since 1989 (1991) in the vision to build up a common database for the printed European cultural heritage which brings together virtually holdings and riches dispersed throughout all of Europe. Beyond its value for cataloguing and retrieval, the database is a highly specialised tool for research allowing a global overview of Europe's printed past. The Thesaurus file is very much in the spirit and methods of this common European database approach, and thus may teach us some methodologically important lessons on authority work. It is under this aspect that I want to consider the CERL Thesaurus.

## 2. Creation and logic of the CERL Thesaurus

2.1. The name of this file, "Thesaurus", is not arbitrary. It points to a very important fact in European cataloguing and creation of authority files. According to our different national cataloguing rules, but maybe beyond that, because of our different cultural traditions and therefore demands or expectations of our different language speaking publics, authority control on national (or regional or local) level leads to different definitions of entry forms (= main heading, = controlled form). A traditionally "universal" place, the holy city of Rome, is known by at least three concurrent name forms: Roma (Italian and Latin), Rome (English and French - by chance), Rom (German and Swedish). And this is as we all know an easy example. Imagine what names you find for this easy place in the tradition of printing (Romae, etc.). There is no uniformity, not in history, not in language, not in cataloguing and not in users' expectations. What authority control achieves is to define (again more or less arbitrarily) one of these forms as standard and list all variants as such in one record for this entity. With this modern instrument of cataloguing in principle the question which form is the correct standard form would be overcome. A powerful implementation of any authority file in a cataloguing or better retrieval system should always allow the search under EVERY form of name contained in the authority record. The standard form remains however relevant, for example for display of the record or display of the form of the name in the bibliographic

record. We, catataloguers, do not seem able to overcome our national expectations and let go of our "preferred form". The Thesaurus is a clever reaction to this fact. The Thesaurus is not prescriptive on the standard form. The CERL Thesaurus takes in all authority records created by any cataloguing agency and maintains the standard form proposed by each agency, indicating that it is the preferred form choosen by this or that agency. The aim is not to be prescriptive in the choice of the standard form. Therefore the format of the CERL Thesaurus can be only inspired by UNIMARC, as UNIMARC only allows one standard form, not several parallel standard forms. But even such an internationally clever policy decision does not overcome all problems. Which form to display first by searching the Thesaurus? The decision was to take the first in alphabet - that is at least something on which cataloguers and users can agree. Yet the results are not always convincing:
It means Rom for Roma, Parigi for Paris, although Firenze for Florence. Display the nationally correct form? Well, even for places and for most entities in Europe this might create other problems for all those who changed names or national borders. There is no satisfactory answer to this question of multi-language and multi-history. The computers are easier to be satisfied: they need a number which can be standardized. I am convinced that standard numbers in the field of authority will become inevitable. In principle, CERL would be the right body to assign a CERL standard number to places, printers and publishers and personal names of early printing.

2.2. The CERL Thesaurus brings together authority records created elsewhere. That is the whole philosophy of CERL: merging in one file originally separate files coming from different and independent projects. That is what CERL does in creating the HPB file. For internationally shared authority work this methodology is interesting, as they experiment elsewhere with search engins, cross file searching and here with physical merging. You will not be surprised that I am quite concerned about the rationality and effectiveness of this undertaking. Yet I think it is feasible for this after all limited field of early printing. The difficulties lie less in merging the original files by application of as much machine power as possible, than in the maintenance and machine updating of the once created new records by updated information coming from the original files. In the actual state of the CERL Thesaurus we can see that machine merging is feasible for entities like place names and we are quite confident on personal names. Still it needs a careful analysis of the original file and a careful mapping and duplication check to the existing file.
For the place name part of the Thesaurus, the file was built up by the file originally created by the Bavarian State Library published in book form. In fact, none of the files actually present in HPB is really linked to an authority file for place names. What most files contain is a (more or less consistently applied) standardized form of the place name and the form found in the book itself. The file was then enhanced by stamdardized forms from the Stockholm file and it is planned to do the same for the forms of St. Petersburg. In both cases, that is not an authority file, but a list of names giving the standardized form and a variant. The machine integration of the Cathedral libraries file and of EDIT 16 place names was less successful, but as it is a relatively small number of records, integration will be done manually. In fact, manual updating may be more successful. Machine merging must be complemented by strong manual editing, and we are not yet in the process of implementing updating routines.
As for printers' and publishers the authority files of The Hague, Paris and Zagreb are already part of the CERL Thesaurus. Machine merging has not yet been undertaken. This part should again be complemented by a file from EDIT 16 and the Cathedral libraries. Here huge problems arise because cataloguing or better the definition of the standardized forms differ very much. Printers are sometimes considered as personal names, sometimes as corporate bodies and this is reflected in the differing structures of the names.
As for persons, the first imput in the CERL Thesaurus is also from the Royal Library in The

Hague and the authority records from ESTC have been integrated, again for the time being without merging. Here the next step will be to take in the relevant parts of the German Personennamendatei (PND) which is highly specialized in names of early printing, containing all the classic and medieval authors and all names coming from the German conversion projects of holdings before 1850.
My guess is that the CERL Thesaurus will very quickly become an independent tool for intellectual editing of authority information on early printing. For building it up, all existing files can be helpful, but then it will lead its own life, which again makes sense in this particularly specialised area of cataloguing - whereas this is no solution for ongoing authority work.

## 3. Function of the CERL Thesaurus relative to HPB

Why did CERL start to build up this Thesaurus file? There are two complementary reasons: one, that lots of files did arrive for the inclusion into HPB which in their original context are based on authority files, mostly for names of persons (and corporate authors), and also printers and publishers. This original link to an authority file is kept in the HPB by maintaining the number for linking, but is of no help in searching as this number lost its linking value. At the same time all information contained in the authority records referred to is lost for HPB. Only the standardized form is maintained which is just one form. The second reason is HPB itself. As I said this is a highly specialised information tool, and the information contained in the records is carefully maintained, mapped and thus retrievable in a way that is most adequate for early printing, much better than what our OPACs can do. Still this sophisticated retrieval cannot overcome by itself the different traditions of cataloguing transported in the records. So the Thesaurus was meant to directly help the user of HPB to find his way through the multiplicity of cataloguing traditions and to give him the best and completest of results for a particular search without asking him to provide himself the whole intellectual work (like knowing all variants of a place name).
For the time being the decision is *not* to implement any new linking structure, now referring to the CERL Thesaurus record, into the HPB. The CERL Thesaurus remains a completely separate database even run on a different system and in another continent. The so called "assisted search" takes the user of the HPB into the CERL Thesaurus. He identifies the record and all (relevant) name forms are transported into the HPB to perform a more complete search. At first hindsight, this is fine and gives better retrieval results. Still I am not convinced that this is the final answer to the search problems in HPB. Some of them cannot be overcome without a firm linking structure.
Three examples taken from the place names may illustrate my doubts:
1. Fictitious place names cannot be included into this assisted searching.
2. Homonymous names create huge problems and falsify the retrieval results.
Frankfurt - Frankfurt am Main, Frankfurt an der Oder
3. Homonymous name variants create the same kind of confusion.

## 4. Function of the CERL Thesaurus as a separate information tool

Even if the functionality of assisted searching is not perfect, it helps. But the true value of the CERL Thesaurus may be found in itself - in a tool that gives us the possibility to bring together and to edit authority information for early printing in a single and authoritative environment and maybe even scientific context. This also allows for unique possibilities of adding value by joining different information tools all concerned with early printing.

There are already first examples. The place name part of the CERL Thesaurus was built up on the basis of a publication done by the Bavarian State Library in 1991 which needs reediting. The reediting is done inside the CERL Thesaurus and we profit from the enhancement we can get through other files, like the one of Cathedral libraries, from Zagreb and place names extracted from the file of Stockholm. We now try to build up systematically this file - as we did for the names of medieval and classic authors. Again it is a limited number of entities to be taken into account and we can do editorial work be using and extracting names from reference sources which do exist and often come from the 19th century (a time where specialised authority work must has also been very much undertaken). Reediting this file means to supplement missing names relevant in the field of early printing, which we did in building up a supplement by extracting information from reference works, containing round about 900 new places. Reediting also means to find alternative name forms, to indicate sources of reference and other relevant information about the place. A particular concern is to give information on fictitious place names. So the place name Thesaurus file can develop into a precious tool for information on places of early printing. One major aim is to supplement this reedited place name file with the information of geographic coordinates contained in the Ghetty Thesaurus file. This will be done by machine procedure once our name file is complete enough to include as many relevant names as possible. With these coordinates we can create an electronic map which would free us from the historically horrible question of country codes and maybe give a better result to research questions like "French imprints of the 16th century".

Geographic historical information is also provided by other projects not limited to the history of printing. Bavaria for instance creates an electronic tool for its regional history. Why not link this information to records in the CERL Thesaurus so that you can switch from one instrument into the next.

One of these adding value methods is already reality inside the CERL Thesaurus. After searching a place name, you may wish to see all printers and publishers working at this place. The tool for that is in place - it is achieved through internal linking between the place name and the printers' and publishers' file. It depends on the fact, that a place name is contained in the imprint name record. Not all incoming files will have this feature, and manual editing will be necessary to provide reliable information.

The imprint name records also allow for further connections. Here we started to explore into the possibility to link into the rich Italian projects of digitization of printers' or publishers' devices. Although this information would not directly lead to the edition in which this device was found (what is done inside the original projects), it would give further information useful in the context of early printing. Last but not least that would be an appropriate place where to host or connect the important information of printing characters, which for the time being is contained in very valuable traditional volumes - which may become digitized and thus serve an old need in a new form.

As for personal names the situation may be different. Here, the national agencies may be much closer to the maintenance of information, but still why not imagine that one of my favourite projects, linking the name records of classical and medieval authors as contained in PAN and PMA and which really belong to all Europe (if not beyond) to the digitized copies of those reference sources quoted in the records. It would very often help to identify the correct authority record for an author. In this area the allocation of a standard number would also be of a huge value, because again, the discussion which entry form is the best, will never end and prescription is not the right method.

Let me conclude. With these last sentences, you can see that I tried to do justice to the invitation I originally received and which was to talk about the Medieval Name authority file, PMA, to which my name is linked at least in the mind of Mauro Guerrini whom I want to thank for this invitation. Talking about the CERL Thesaurus gave me the opportunity to talk

about a work in progress to which I am still related and to talk about reuse, reorganisation, adding value to existing work on authority control. I firmly believe that authority work is the way to better disclose the richness of the bibliographic universe to those who need our services. We must do this in a sensible way, joining our forces, making the best of what exists and learning from the experiences of each other. For this, this conference gives us an ideal opportunity which I enjoy very much.