Making all Swedish out-of copyright literary books available for mining: a digitization project in development

Mats Malm

Literary fiction, including drama, is an invaluable source for understanding the movements of society. It is not only a source for studies on literature and art: it has the ability to demonstrate what forces and ideas are at work in the society of its time. The issues may concern social order and the structure of society, world view, gender, national identity, the foreign, ethnic or sexual minorities etc., alterations in the meaning of words and concepts, or they may be purely aesthetic. The examples of possible scholarly approaches are many more: fiction is the space of constant negotiations between old values and new ideas. The novel offers lavish room to shape and portray problems and oppositions of its time: the short story and drama offer the means to more succinctly focus on a specific issue. Literary fiction, when digitized, offers an arena for intense interdisciplinary exchange and development of methods and theories.

While the literary works in many language areas have to a large extent been digitized by now, Sweden stands out for its lack of digitized literary cultural heritage. There are initiatives such as the Swedish Literature Bank (http://litteraturbanken.se), Dramawebben (http://dramawebben.se) and Project Runeberg (http://runeberg.org), but these are comparatively small and not the products of a national, synchronized effort at digitizing. National funding has been distributed for separate, minor projects, but mainly as labour market measures, in a way that has not enabled structured digitization and accessibility on a larger scale. Nor has the Royal Library received funding to cover the demands posed.

This presentation sketches a digitization effort that aims to solve to this problem, starting with digitizing and making accessible for large scale analysis first all literary books authored in Swedish, then literary journals, until about 1900 or later, when copy-right can be handled. If it proves a viable way, it can also be reproduced in other places, covering more and more of the total of Swedish books.

In a collaboration between The Swedish Literature Bank, Gothenburg University Library and The Swedish Migration Centre, a plan for creating a digitization workshop of 20 persons during 5 years is being made. The plan depends on the

participation of the city of Gothenburg and funding of equipment, but as it means no real cost to the city and the equipment part is comparatively small, the prognosis is good. Most of the funding will be made by governmental funding through The Swedish Migration Centre. The crew will be picked out from unemployed persons with a degree of disability; the tasks will be adapted so as to optimize effectivity. The workshop will be established elsewhere, but the production flow will be integrated in the University Library's production line. Every book page will be made available as a PDF file with integrated OCR by the Library, and metadata both for the Library, the national catalogue Libris and for The Swedish Literature Bank will be created. A selected amount of old books will be transcribed and chosen works will be proof-read for greater accuracy, but the material on the whole will consist in OCR-ed books with basic metadata. The whole material will be collected and made available for research in much more adjusted ways in The Swedish Literature Bank: here, complex searches will be possible to make along with machinery for topic modeling etc. The Swedish Language Bank is developed on the infrastructure of The Swedish Language Bank, which means that much of the technological solutions of the Language Bank will be available through the Literature Bank and that the whole material will also be available as one of the corpora of the Language Bank. Thus, all techniques offered by the Language Bank will be possible to impose on the material as well – the Literature Bank is directed toward a wider set of humanistic researchers, students, as well as the public.

The selection of Swedish books (all editions), i.e. authored in Swedish, is very roughly expected to cover 20.000 volumes: the expectation is that after that, there will be room for journals and humanistic publications within, not least, history. Personnel from the University Library and the Literature Bank will instruct and supervise the process. As The Swedish Migration Centre establishes workshops all over the country, our ambition is that the plan shall be emulated at other university libraries, directed toward other selections, to the effect that a very large portion of Swedish books can be digitized within the next six years, starting in August 2015.

If the project concept proves successful, it should be possible to duplicate it in order for other university libraries to cover other selections of Swedish out-of-copyright books. In six years, there will then be a wealth of material which can provide the basis of a number of innovative interdisciplinary projects.

The plan is currently being worked out in its details. The paper will report of the status of the project as it has proceeded by the time of the workshop in November.