Today, many historic serial sources are available only in print format, i.e. as printed books, sheets, loose-leaf-collections and other kinds of paper-based documents. Unfortunately, the conversion of a document from printed to structured digital format (i.e., transcribing, tagging, encoding, and indexing; “in-depth-indexing” in the remainder of this paper) is a very time-consuming and cost-intensive task. Because of this high effort, archives usually don't perform in-depth-indexing of archived materials; in most cases, sources are enriched with some formal metadata (provenience, publication information, etc.) only.

In this contribution we show a) how the tedious work of in-depth-indexing can be simplified by technology and b) how the workload can be distributed to a large number of people via crowdsourcing. We describe an in-depth-indexing project where approx. 700 voluntary contributors transcribed the German WW1 casualty lists, which comprise 31,000 newspaper-format pages displaying 8.5 million entries.

In cooperation with the German “Verein für Computergenealogie” (Society for Computer Genealogy), we built a novel crowdsourcing platform: CG-DES. CG-DES has an easy-to-use browser-based interface that heavily contributed to the motivation of the volunteers.

We set up seven hypotheses how volunteers behave and underpin these hypotheses with statistical data collected during CG-DES operation time:

1. 80% of the data is entered by 20% of the volunteers. (Pareto principle)
2. Most of the work is in the evening hours and on weekends.
3. “Power users” leave the project less than other volunteers.
4. After a volunteer joins the project the number of edits per day increases strongly, then drops to a long-term average value.
5. The longer a volunteer works on the project the less errors he/she makes.
6. “Power users” make less errors.
7. Volunteers that leave the project after a very short time contribute only little and cause a lot of work.

As it is very likely that these hypotheses are valid for other crowdsourcing projects, it is very important to explore the above given hypotheses. It can be assumed that our project yields valid figures because of the large project workload, the long project runtime, and the large number of contributors.