Dr. Birgit Welke
Charité - Universitätsmedizin Berlin
Dr. Bernhard Krause
QLEO Science GmbH

Automatically generated Profiles for Experts, Institutions and Working Groups

The current research information system (CRIS) system at the Charité - Universitätsmedizin Berlin has been in use for more than 20 years. Since 2001, the system has also been used to calculate the publication-related performance for the performance-oriented allocation of funds. Due to this it contains a wealth of confirmed information that makes it possible to automatically identify experts, generate research profiles and identify networks and collaborations both within and outside the Charité. This is done by comparing research metadata with the Medical Subject Headings (MeSH) thesaurus. Cooperating institutions outside the Charité are identified by their ROR identifiers. The automatic generation of research profiles for individuals, clinics and institutes in our research information system can support strategic corporate development in identifying new research focal points and expanding and specifically promoting cooperation networks.

Research information systems are used to collect research meta data in order to present and evaluate the output of experts, institutions and working groups. Master data can often be imported via interfaces to the company's existing HR and ERP systems.

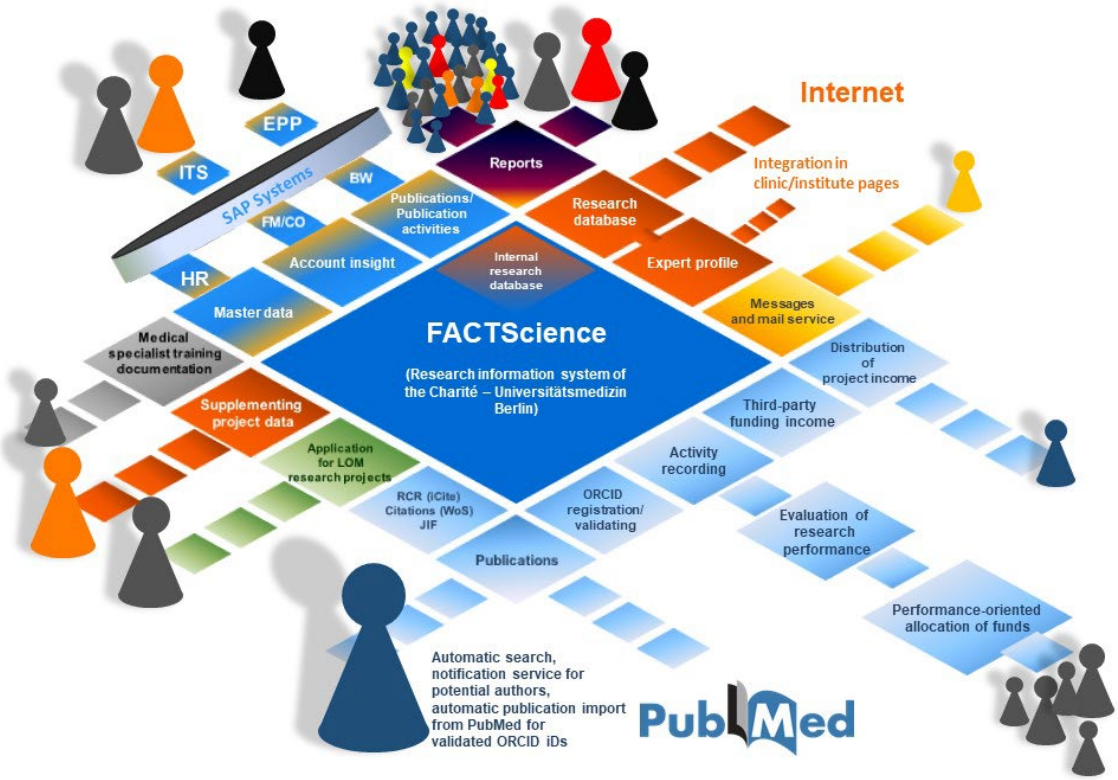

Figure 1: CRIS System at the Charité - Universitätsmedizin Berlin

Publication meta databases are used to deliver consolidated data sets.[1]
In addition, this data can be enriched by

- either calculating the journal impact factor or other bibliometric parameters[2]
- identifying researchers or research organizations by their persistent identifiers (e.g. ORCID or ROR) in order to display collaboration networks and maps[3].

However, these findings usually remain within the walls of the organizations studied.
In biomedical and clinical sciences we can use the Medical Subject Headings (MeSH) thesaurus used for indexing articles in PubMed[4]. MeSH has proved its worth in identifying the research priorities of experts, as published in the Charité research database for researchers who make their data public. But how is this calculated?

MeSH descriptors are organized in 16 categories: category A for anatomic terms, category B for organisms, C for diseases, D for drugs and chemicals, etc. Each category is further divided into subcategories. Within each subcategory, descriptors are arrayed hierarchically from most general to most specific in up to thirteen hierarchical levels. Because of the branching structure of the hierarchies, these lists are sometimes referred to as "trees". Each MeSH descriptor appears in at least one place in the tree, and may appear in as many additional places as may be appropriate. The lower a descriptor is in the hierarchy, the more specific it is:

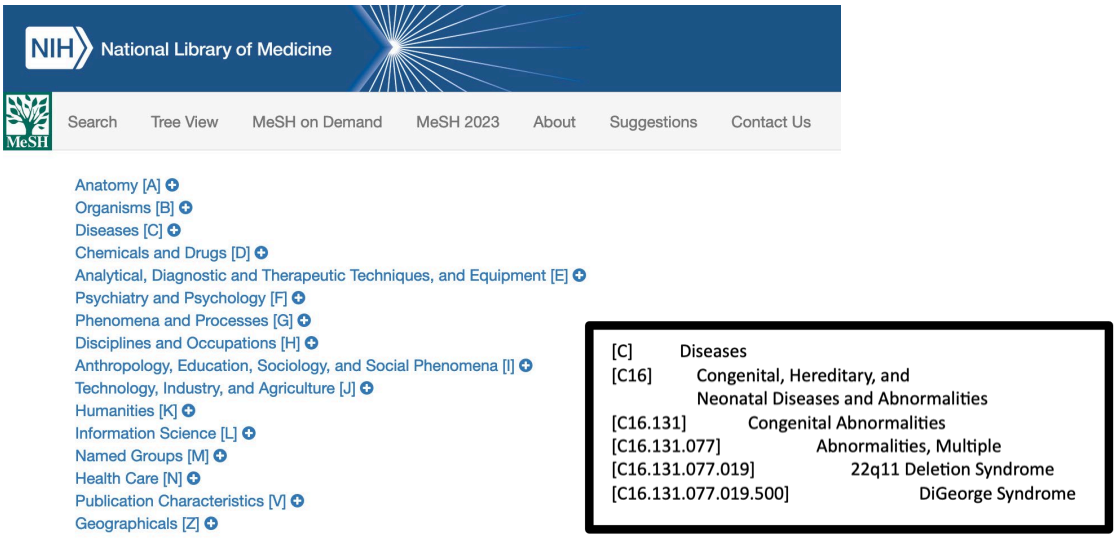

Figure 2 Structure of hierarchical MeSH Headings [5]

The MeSH descriptors and their entry terms are now matched with CRIS meta data in this system with more than 150000 publication and 30000 project data sets. Findings in e.g. title, abstract or keywords are valued differently. The lower a term is located in the mesh tree, the higher the find is rated in the metadata.

The frequency of occurrence is also evaluated. It is also possible to include the year of publication in the assessment, e.g. the younger a publication the higher the value of the find.

In the first step, the result is a scored list of keywords that can be calculated for each publication, project or other research metadata set. In the second step, we can aggregate the evaluated keywords not only for individual persons or the entire Charité but also for smaller institutional units such as clinics, institutes, departments and working groups.

[1] *e.g. PubMed, OpenAlex, ORCID, Unpaywall, CrossRef and others behind paywalls*
[2] *e.g. Journal Citation Reports, iCite*
[3] *like SciVal Collaboration Module (https://www.elsevier.com/products/scival/collaboration) or Dimensions (https://www.dimensions.ai)*
[4] *https://www.nlm.nih.gov/mesh/meshhome.html*
[5] *https://meshb.nlm.nih.gov/treeView*

Figure 3: Experts word cloud and co-authoring network

Most CRIS systems present research content as searchable databases and or as lists in Website integrations. This enables the implementation of the required presentation and reporting obligations and - when automatically updated with ongoing metadata - is a great success. We have supplemented these views with researcher profiles, institution profiles and aggregated bibliometric indicators, enriched by co-author networks and overviews of bibliometric indicators and statistics, such as the proportion of open access publications.

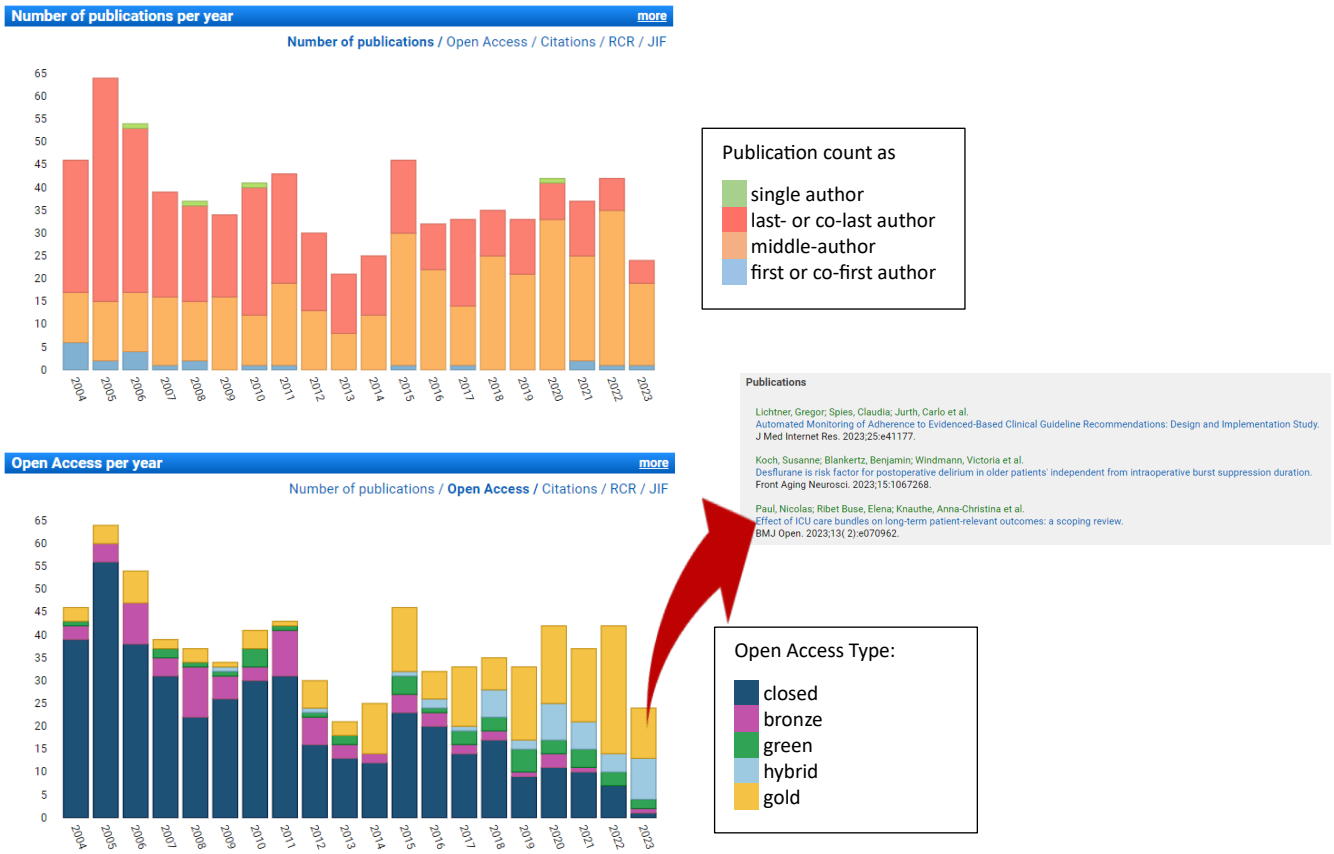

Figure 4: Example for a bibliographic overview

Most of the functionalities already mentioned have been established in Charité's CRIS for many years. The beginnings of the system date back to 1998, when it was originally developed for the performance-oriented allocation of funds at the Charité (EuroCRIS was founded in 2002). A great deal of effort is put into updating the database promptly. A weekly service informs scientists when new publications are listed in PubMed. The ORCID integration in our CRIS enables an automated assignment of publications to researchers and thus also to the associated clinics and institutes. 93 % of our researchers who publish have an ORCID ID registered or validated via our CRIS.

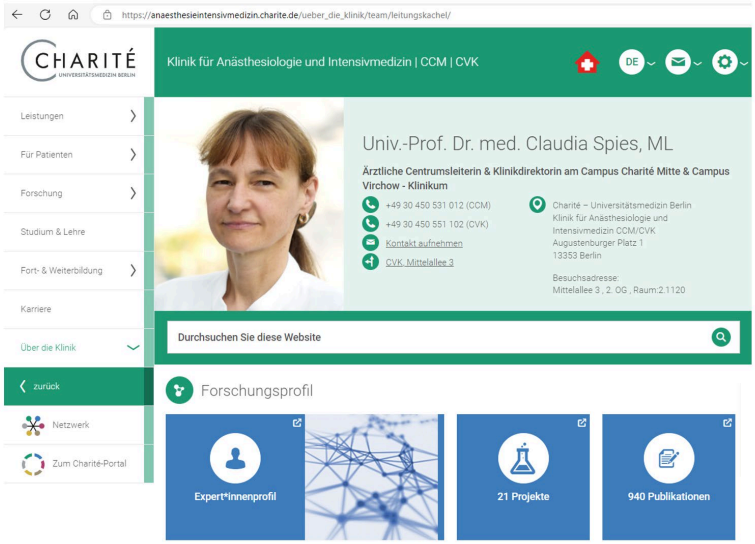

Figure 5: Integration of CRIS in the CMS

The system output is now integrated into the current CMS presentation in more than thousand places with dynamically updated information and profiles. The database is also used to fulfill the requirements of the core data set research (Kerndatensatz Forschung[6]), e.g. the assignment of scientists to publications, projects and third-party funds and cooperating institutions. Aggregated information is used in SAP BI management functions.

In the next step, we import the GRID/ROR IDs for the affiliation details in the publication metadata. This allows us to create the list of cooperating research organizations from the perspective of the entire research organization. This results in the data structure shown in Figure 6.

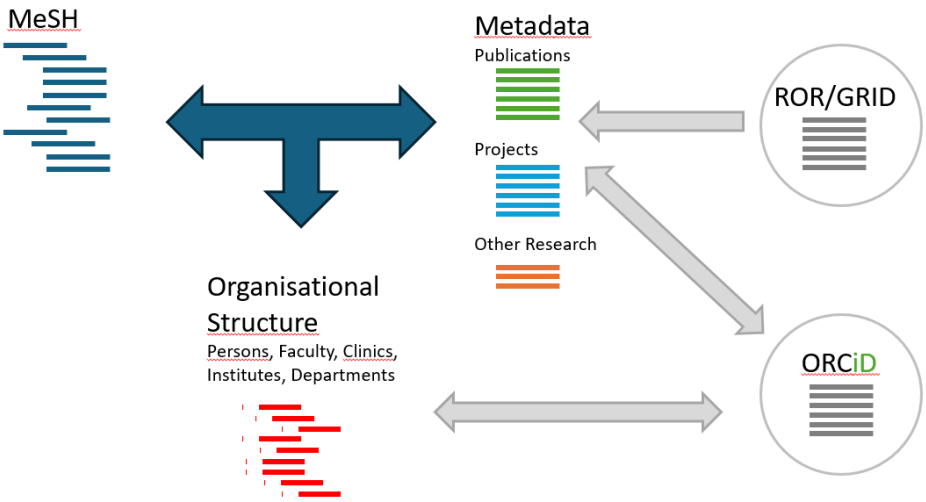

Figure 6: Overview of data structure for analyzing institutional research profiles

[6] https://kerndatensatz-forschung.de/

This is of course also done by various providers who offer detailed assessments and research profiles for research organizations.

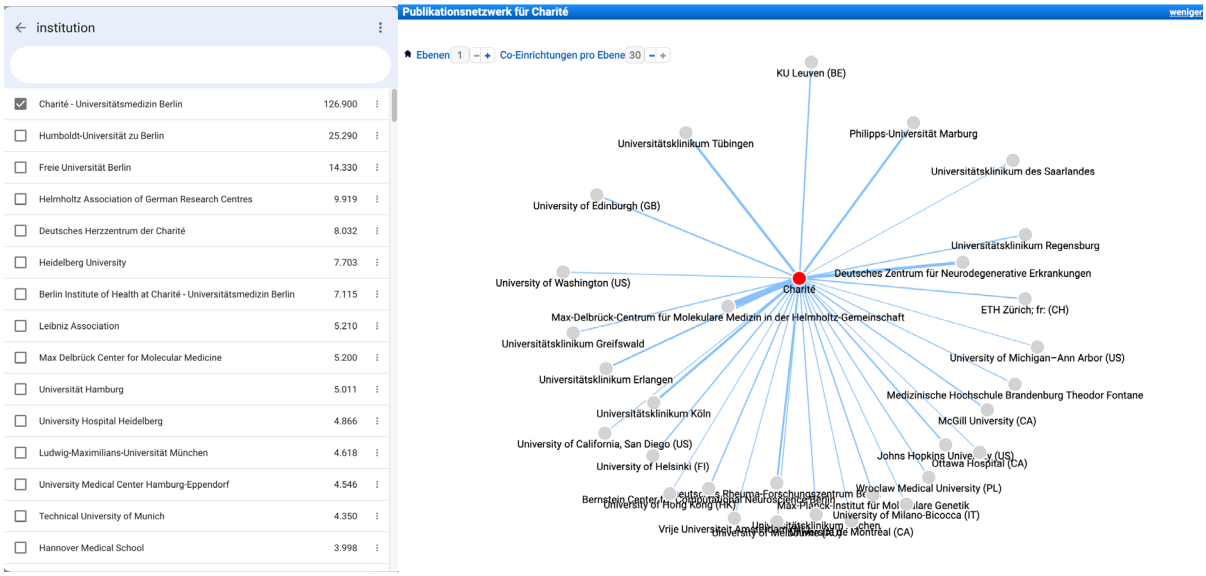

Figure 7: Collaborations derived from institutional publication data

In addition, we can also apply the evaluations to the internal organizational structure at CRIS. This shows which centers, clinics, institutes, departments and working groups collaborate with which external research organizations. By extending the network not only to direct collaborations, but also showing the complexity of existing collaborations in a 2-level view, we obtain more information.

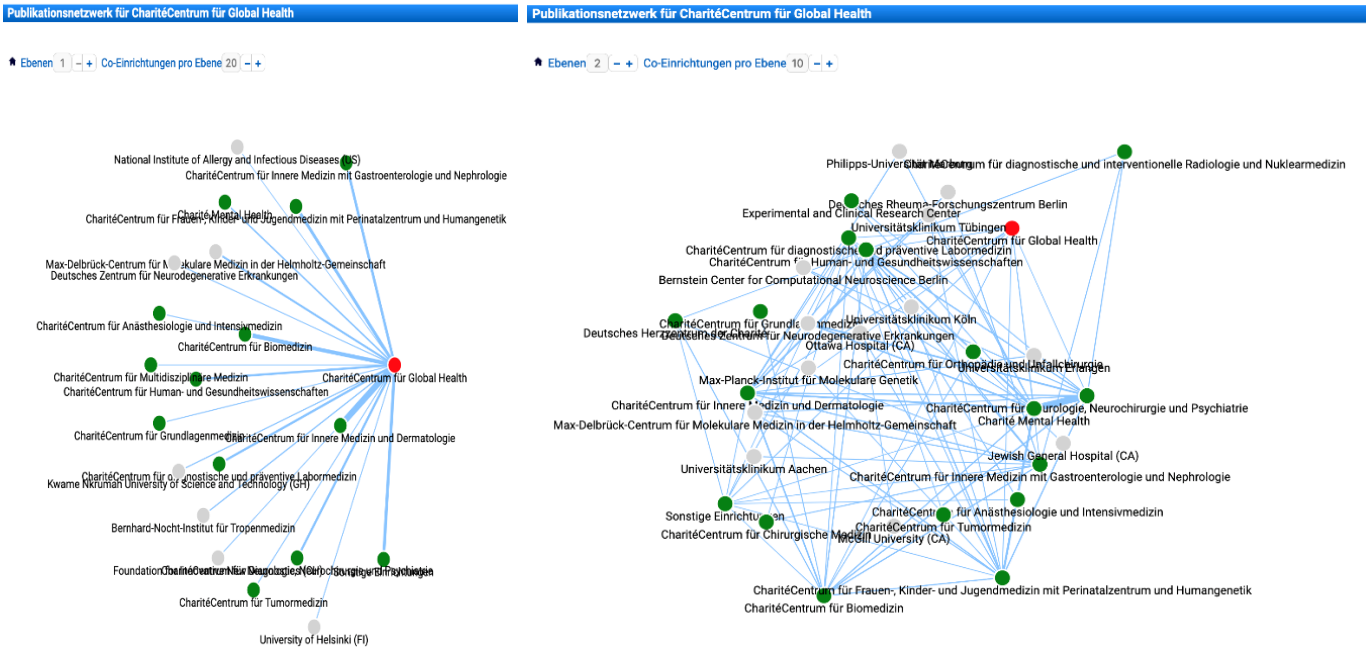

Figure 8: Internal and external collaborations provide insights into large research institutions

Collaborations can be broken down by research areas and topics. Here, too, we use the MeSH Thesaurus to identify the most important research topics at the Charité that produce research results.

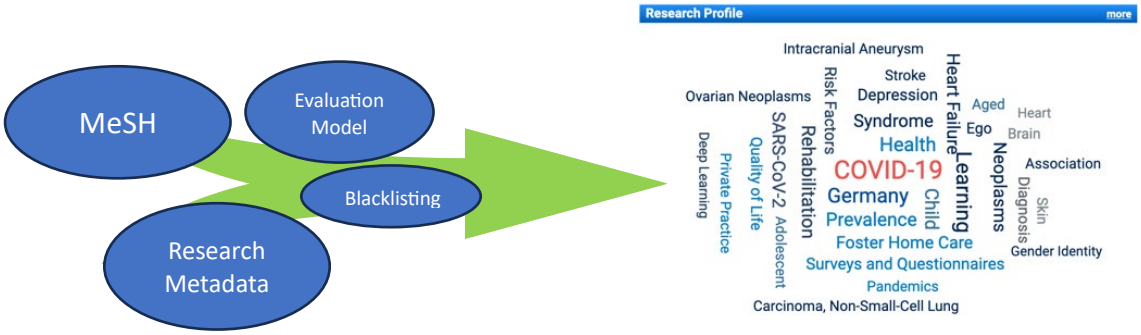

Figure 9: Main research topics resulting from research metadata

It is no surprise, that CoVid19 dominated most other research areas in 2020 and 2021. But who were the key players collaborating with our organization in this area?

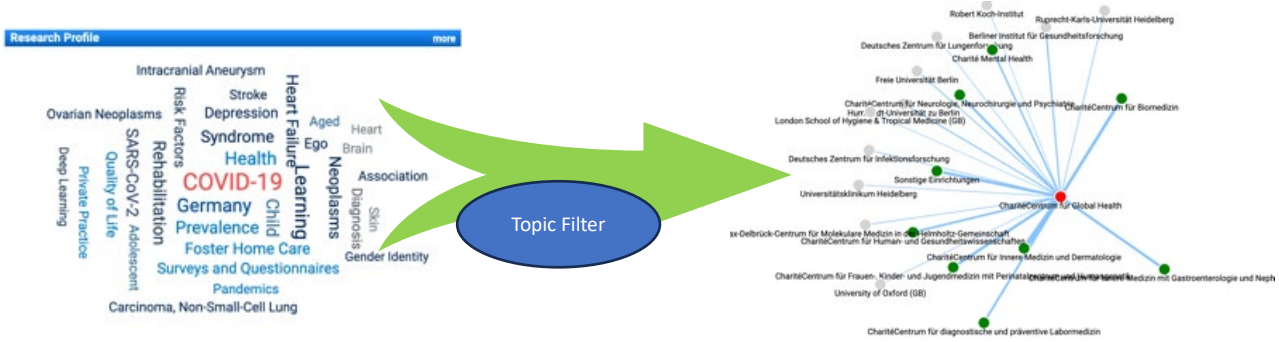

Figure 10: Collaboration network on specific topics

These are just examples. Linking the data from all the sources mentioned can give experts insights into more detailed facts and now allows qualified and quantified answers to more of the following questions:

- What are the current research areas of the research groups?
- How does the focus change over time?
- Which research group has focused on a specific research area?

It took a long time to introduce and establish the functions presented here in a research institution as large as the Charité - Universitätsmedizin Berlin. The new concepts and enhancements presented here will supplement the already published components of Charité's CRIS in the coming months.