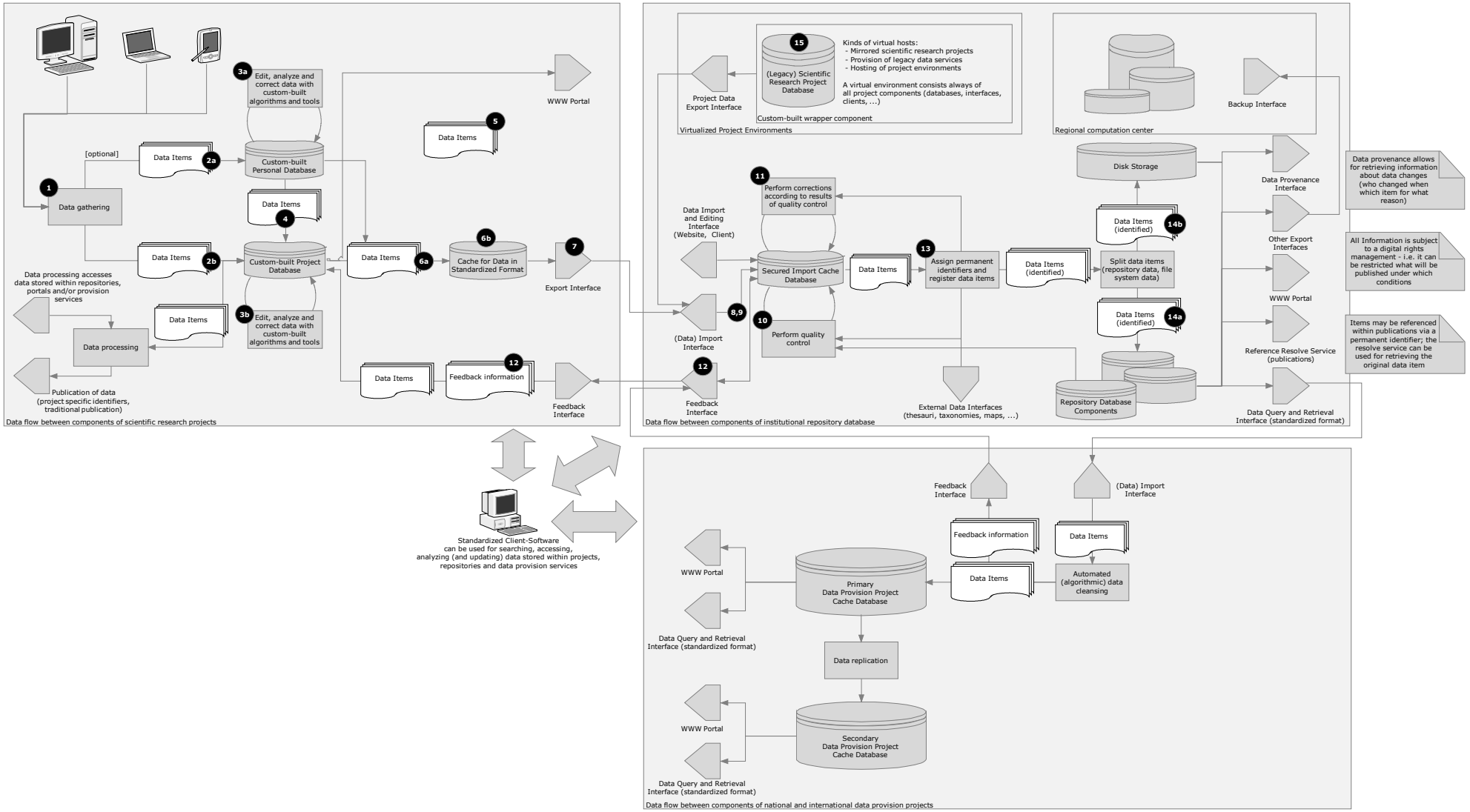
Proposal: Institutional Repositories for Biodiversity and Ecology Data

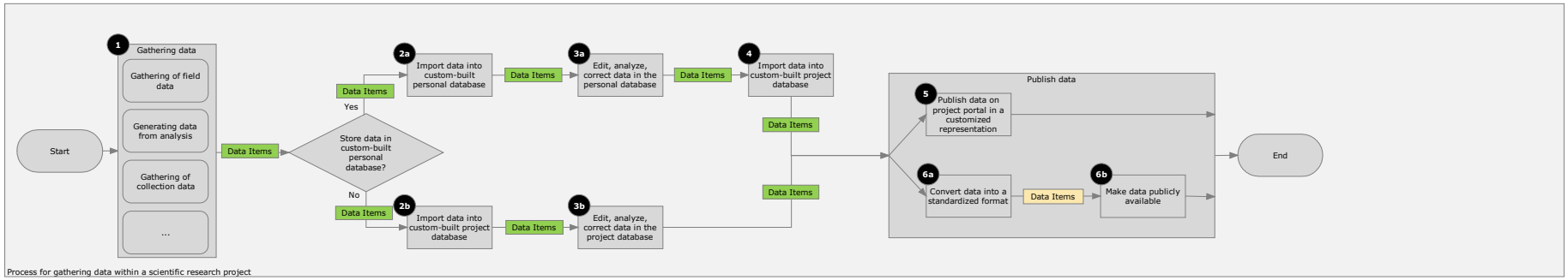
Central Laboratory for DNA Analytics and Ecoinformatics, Chair for Applied Computer Science IV, Chair for Applied Computer Science I
University of Bayreuth, Bayreuth, Germany

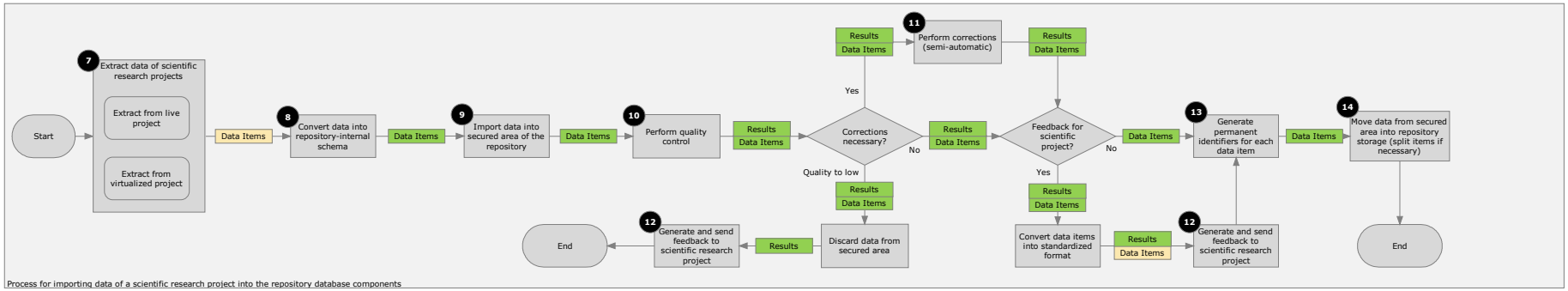
IT-Center of the Staatliche Naturwissenschaftliche Sammlungen Bayerns (SNSB)
Munich, Germany

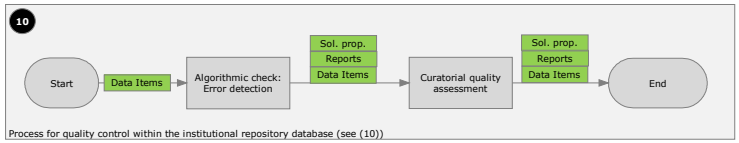
Overview of (logical and physical) components and flow of data



Description of important activities







The concept is divided into two parts; the first part on page 1 shows main excerpts of the data flow within the components of scientific research projects, institutional repositories and data provision projects. The second part on page 2 shows how normal procedures are executed within the group boxes of the first part modeled as processes. Processes are advantageous for the description of complex application scenarios since they provide a good means for documenting, modularizing and making single complex application cases transparent (processes thus contribute to the quality of software). Furthermore they can be seen as a part of a software specification that tells what has to be done by whom in which order with what data and with which tools in an easy-to-understand manner. Parts of both figures are interlinked with each other via white numbers shown in black circles.

The overall data flow starts with - in general - gathering of data (1) which is stored within a personal database or a project database. Both databases (can) use a custom-built schema which is specially tailored to the requirements of an application/project domain. In case data items are stored within a personal database (2a) they might be later on transferred to the project database (4). Both databases can be accessed with custom-built client software in order to perform data analysis and corrections or simply edit single data items (3a, 3b). Special data processing steps can be considered that also use data from external sources and which also provide the ability to publish results (in case of data publication, data items carry a project-specific identifier). After data arrived in the project database it can be distributed in two ways; first, it can be published on the projects website with a custom or standardized format (5). Second, data can be transferred to an Institutional Repository Database. Since many projects can be connected to such a repository, it is mandatory to convert data from the custom-built schema of the project database into a standardized representation (6a). In order to increase performance during the later on extraction, data can be thought of being cashed within a special buffer database (6b) after the conversion

A special data import mechanism then collects data from the buffer database via dedicated interfaces (7) and transfers data items into a secured caching area of the repository (9) after performing a data conversion from the standardized exchange format into the internal, optimized repository representation (8). After the import into the secured area, a quality control (10) can be performed in a semi-automatic manner; first, an algorithmic verification process can be run that can detects first inconsistencies; the results of this automatic error detection can then be used in a following manual assessment of the imported data items. This assessment then finally yields results that can be used in the following error correction phase (11) which can be performed semi-automatically. In case data quality is too low, data might be discarded from the repository. In both cases the corresponding data producer (scientific research project) will be informed about the findings of the quality assessment and possible corrections are fed back into the project (12).

After passing the quality assessment, data are moved from the secured area into the repository databases. Therefore, data are assigned permanent identifiers which identify a single item uniquely starting from this point on (13). The repository itself is split into several components which greatly depend on the purpose of such a repository (i.e. what data are to be stored) and even might contain normal file stores, e.g. for saving single entities such as videos, images or audio recordings. Thus data items are maybe split into several chunks in order to fit in the repository (14a, 14b). After data has been moved into the repository, it is available via several export interfaces. Additional information about the each action that manipulated data can be retrieved via the Data Provenance Interface of the repository.

In case the owners of data agreed with a publication of their items, they are made accessible to data provision projects via standardized interfaces and wrapper components. Feedback of these projects can be sent to the repository and will undergo the same quality assessment process as made outlined above. Thus, the two-staged architecture of the repository ensures that only qualified data are made available - the free manipulation of data within the repository is already conceptually not possible since no direct way between research projects and provision projects with the core repository database exists.

Legacy projects (projects whose funding is running out, whose technical bases is outdated or projects that plan to move data soon to the repository) can be hosted within virtualized environments (15). Please note that it is also possible to host a complete project environment within such a virtual machine for projects that do not have a suitable infrastructure at their disposal. Data of such environments can then be also transferred into the repository.