Development of a Large-scale Neuroimages and Clinical Variables Data Atlas in the neuGRID4You (N4U) project

Kamran Munir·, Khawar Hasham Ahmad, Richard McClatchey

Centre for Complex Cooperative Systems (CCCS), Department of Computer Science and Creative Technologies (CSCT), University of the West of England (UWE), Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, United Kingdom.

Abstract.

Exceptional growth in the availability of large-scale clinical imaging datasets has led to the development of computational infrastructures that offer scientists access to image repositories and associated clinical variables data. The EU FP7 neuGRID and its follow on neuGRID4You (N4U) projects provide a leading e-Infrastructure where neuroscientists can find core services and resources for brain image analysis. The core component of this e-Infrastructure is the N4U Virtual Laboratory, which offers easy access for neuroscientists to a wide range of datasets and algorithms, pipelines, computational resources, services, and associated support services. The foundation of this virtual laboratory is a massive data store plus a set of information services collectively called the 'Data Atlas'. This data atlas stores datasets, clinical study data, data dictionaries, algorithm/pipeline definitions, and provides interfaces for parameterised querying so that neuroscientists can perform analyses on required datasets. This paper presents the overall design and development of the Data Atlas, its associated dataset indexing and retrieval services that originated from the development of the N4U Virtual Laboratory in the EU FP7 N4U project in the light of detailed user requirements.

Keywords:

Data atlas, neuroscience; data integration, data analysis; information retrieval

1. Introduction

With the enormous increase in variety and size of clinical datasets, and increasing information complexity in biomedical research, biomedical researchers are often faced with severe difficulties in data management and information retrieval. The neuGRID4You project (N4U Project, grant agreement n. 283562, 2011-2014) provides an e-Infrastructure where neuroscientists can find services and resources for neuroimaging analyses [1]. It provides an e-Science environment through a Virtual Laboratory (whose model is depicted in Figure 1) that offers neuroscientists access to a wide range of clinical and neuroimaging datasets, algorithm applications, computational resources, services, and information support. This virtual laboratory has been mainly developed for neuroscientists but is also applicable and adaptable to other user communities. The foundation of the N4U Virtual Laboratory is a massive data store, named the Analysis Base, and a set of Information Services. The combination of the Analysis Base and its Information Services is termed the 'Data Atlas'. The Analysis Base is a structured data store that stores datasets, clinical study data, data dictionaries, and algorithm/pipeline definitions. The Analysis Base is accessed through services that store or index datasets and algorithm pipeline

· This is the corresponding author
Email address: kamran2.munir@uwe.ac.uk (Kamran Munir)

Journal of Biomedical Informatics

definitions so that users can perform any investigation or data analysis through the N4U Virtual Laboratory. This subset of Information Services is termed the Persistency Service. Moreover, the Information Services also provide various interfaces for the parameterised querying of datasets to assist virtual laboratory users (such as neuroscientists) in defining and executing their analyses on filtered datasets. This subset of Information Services is termed the Querying Service. The outcome generated by the Querying Service can be exported in various formats, such as XML and CSV, which is then used in other software applications to generate or to perform an analysis, e.g. by using the CRISTAL software [2].

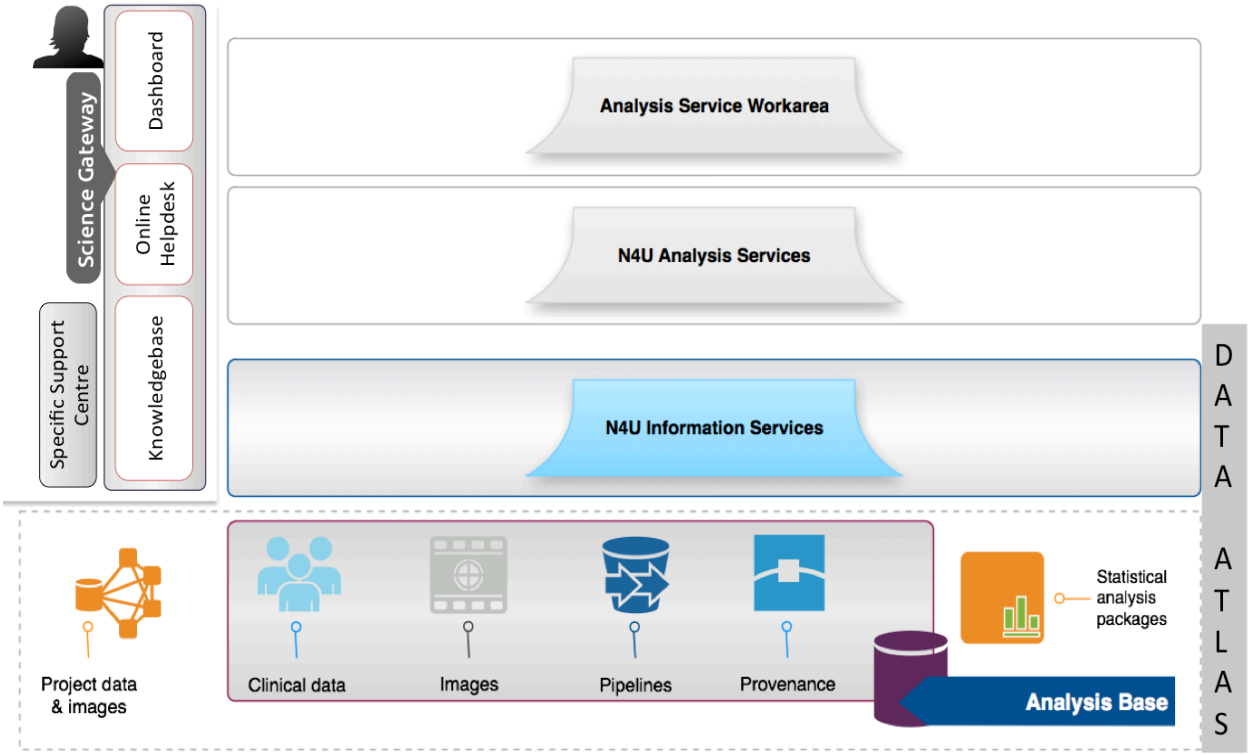


Figure 1: The N4U Virtual Laboratory architecture.

The design and development of the Analysis Base and its Information Services (i.e. both Persistency and Querying Services) has been carried out in the light of detailed N4U users requirements [3], which have been analysed during the requirements gathering and specification phases of the N4U project. Moreover, there have been an enormous number of evolving requirements, such as the increased complexity and heterogeneity of datasets, which has led to further refinements in the design specification and implementation of the Analysis Base and its Information Services. The main challenges faced in providing these services relate to, for example: (a) the different formats, structures and semantics of datasets; (b) the fact that all datasets had a large diversity in the number and types of clinical study parameters; (c) the manner in which the original relationships between clinical study parameters and image files (such as brain scans) were maintained also varied across datasets and dataset providers; (d) the fact that comprehensive data dictionaries had to be maintained for each dataset; and (e) that each dataset could have unlimited numbers of clinical variables that could hold different types of values (called clinical variables scores). This paper presents the overall design and development of the Data Atlas, including the development methodologies and algorithms used by its associated information services in the light of the N4U user requirements.

This paper has been organised as follows: Section 2 reviews the relevant approaches in neuroinformatics that aid in managing biomedical data and assist researchers in performing data

analysis. In Section 3, the requirements, design and development of the N4U Analysis Base, which is the foundation of the N4U Virtual Laboratory, are presented. Section 4 outlines the requirements and associated challenges of dealing with large datasets, followed by an introduction to the N4U Information Services that carry out all the required functions of importing, indexing and querying the datasets including, for example, clinical study data, neuroimages, pipelines and algorithms in the Analysis Base. Section 5 presents the indexing of different dataset use cases (examples), pipelines and algorithms, which shows the availability of different types of datasets in N4U and also their internal structures – representing the medical data along with data dictionaries and image files. Section 6 discusses the implemented data retrieval mechanisms, which enable the N4U end-users to perform their queries on clinical data and images in order to retrieve desired data related to pipelines/algorithms. This section is concluded by a description of the use of rich clinical data dictionaries to empower end-users in defining their search criteria. Finally, the outcomes of this research and the identified future research challenges are presented in Section 7.

2. Related Work

Managing huge volumes of data produced from neuroscience research and then enabling their querying has been a major research challenge in the neuroscience community [4]. There has been much work carried out to support the storing and retrieval of neuroscience data. Such efforts can be divided into two main categories; (a) the cataloguing of data concerning neuroscience resources including papers and (b) the management of actual neuroscience experimental data (including neuroimages and clinical variables). These two categories are discussed in the following subsections. This work focuses on providing a mechanism to enable the indexing and retrieval of neuroimages stored on the Grid infrastructure of N4U, which itself is an extension of the earlier neuGRID project (http://www.neugrid.eu).

(a) Data of Neuroscience Resources

In the first category, paper [5] discusses projects that have been developed to discover and integrate neuroscience resources for the use of the neuroscience community. The primary aim of these projects is to provide a single platform for neuroscience researchers to search for the required data based on domain specific keywords. The platform also acts as a search engine so that neuroscientists can discover relevant articles and databases to locate selected data. The Neuroscience Database Gateway (NDG) (http://ndg.sfn.org/) is a Web-based catalogue of neuroscience databases. It provides a registry of neuroscience databases annotated with controlled keywords and supports over 200 databases spanning different neuroscience subdomains, such as the neurophysiology SenseLab [6]. The Neuroscience Information Framework (NIF) [7] provides a “one-stop-shop” for neuroscience researchers to access heterogeneous information resources. It does so by providing three registries: (i) a NIF resource registry, (ii) a NIF database mediator and (iii) a NIF document archiver. NIF is an ontology driven system with at its heart the NIFSTD (or NIF Standard Ontology), which is constructed by integrating various other domain ontologies. DISCO [8] is a web based application developed for NIF to provide features such as resource integration and data searching powered by the NIF backend architecture. It provides concept-driven querying support for data discovery.

The KIND (Knowledge-Based Integration of Neuroscience Data) approach [9] overcomes the problem of data integration of datasets that differ in formats. In KIND, data integration is performed over biological study data that come from different sources such as NTRANS (neuro-transmission database) and CAPROT (calcium-binding protein databases). In order to perform data integration, it uses domain knowledge that describes rules of the domain to bridge the gap between disparate data sources. It then applies the deductive object-oriented language *F-Logic* to support complex data integration. Unlike the N4U Data Atlas, KIND does not provide a user-friendly front-end interface that would allow a user to

build and refine its search queries. Similarly, Entrez Neuron [10] provides a keyword-based search against a coherent repository described in Web Ontology Language (OWL) (http://www.w3.org/2001/sw/wiki/OWL) ontologies such as NeuroDB, ModelDB (from SenseLab) and Subcellular Anatomy Ontology (SAO) [11]. These ontologies are then stored in the Oracle 11g database (http://www.oracle.com) that has built-in support for OWL storage and querying. Unlike Entrez Neuron, the N4U Data Atlas does not at the time of writing support OWL based data sources because data exported to the N4U Data Atlas is not represented in the OWL format. Furthermore, the querying interface of Entrez Neuron is very basic when compared to the N4U Data Atlas, which provides a rich and dynamic querying building interface (discussed in Section 6) for users.

The NeuroLOG [12] project proposed a federated approach to integrating neuroimage files, associated metadata and neuroscience semantic data distributed across disparate sites. A Data Management Layer (DML) has been devised to hide the underlying complexity of data stored in different formats at different sites. In doing so, an additional NeuroLOG database, structured according to defined ontologies, is deployed on each site to achieve complete autonomy of the sites and to facilitate existing data formats stored on the sites. The main focus in NeuroLOG is on the storage and semantic retrieval of the data stored across the sites; however, this approach does not provide provenance information of the neuroimaging analyses conducted over the neuroimages. In comparison, the N4U Data Atlas provides support for data indexing and user-driven query building, and the provenance information of user analyses can be retrieved through the CRISTAL software.

The work presented in this paper is different from the aforementioned efforts in a number of ways. Firstly, the N4U Data Atlas is designed to provide data indexing and querying services to search and locate actual datasets of (brain scan) images and their associated clinical variables and metadata. This information is crucial to support analyses executed by the Analysis Services on the N4U infrastructure using CRISTAL (see [13]). Thus, it is essential to provide a mechanism that enables researchers to discover and access neuroimages stored on the Grid infrastructure, namely, the N4U Grid infrastructure. Secondly, the aforementioned projects dealt with heterogeneous neuroscience resources, but they have not explored the heterogeneity within the neuro-data collected at various neuroscience research labs or centres. As discussed in Section 4.1 of this paper, in N4U each dataset is different from the others, hence providing a uniform indexing and querying mechanism becomes a challenging task. Moreover, the aforementioned projects use ontologies or predefined keywords, which are generated as a result of domain knowledge described in the OWL ontologies, in order to provide keyword- or concept- based searches. However, we have presented a dynamic query-building tool that presents all the clinical variables and their metadata to a user and then allows the user to build a query with dynamic values to obtain the desired result set. In this way, it gives more flexibility to the user and is therefore more customizable. Furthermore, most of the existing work is based on semantics, which is not the focus of this paper. The Data Atlas could however be extended in future by integrating a semantic layer to provide this feature.

(b) Neuroscience Experimental Data

Extensive research has been undertaken in order to support the storage and querying of clinical study data. The Functional Bioinformatics Research Network (FBIRN) [14] presented an extensible data management system for clinical neuroimaging studies. This includes the development of a distributed network infrastructure to support the creation of a federated database consisting of a large sample of neuroimaging datasets. FBIRN also incorporates the provenance information of an analysis. However, this provenance information does not provide an insight regarding the execution times of an analysis. Furthermore, its federated, multi-site data organization approach requires the configuration and deployment of the proposed framework at each site, with a consequent increases in overall complexity. In contrast, our approach is simpler since it provides interfaces for external data sources that can be used to index external data according to a single schema, thus providing a uniform storage view to the

users and external data sources. In N4U all the transformation and complexity is hidden from the external data sources and is handled internally by the Persistency Service (as detailed in Section 5). In doing so, a standard data format has been devised for data providers to import their data into the N4U Analysis Base.

The CNARI framework [15], used at the Human Neuroscience Laboratory at the University of Chicago, presents a database-driven architecture that combines databases for storing fMRI data and workflows for analysis purposes. It proposes the use of relational databases instead of a traditional file-based approach for storing and querying fMRI data [16]. It creates separate databases for each experiment and several tables to store images or user-specific data, which requires a huge storage space. In order to execute the workflows, CNARI employs SWIFT as a workflow engine [17] and exploits its provenance tracking capabilities to maintain provenance information for reproducing an analysis. Unlike CNARI, our proposed schema stores references to the image files available on the N4U Grid infrastructure, thus making it storage efficient. Moreover, it also provides runtime provenance information of each analysis through CRISTAL.

The Human Imaging Database (HID) [18] is an open-source and extensible database schema implemented over the relational databases Oracle (http://www.oracle.com) and PostgreSQL (http://postgresql.org). HID has been designed to operate in a federated environment in which each site has its own HID instance. The images and derived data, which are physically stored on the Grid, are linked back to the HID system located at the site that originally imported the data into the system. Since HID operates in a federated environment, it provides a data integration engine so that the data stored on all sites can be exposed and queried as a single database.

The Neuroinformatics Database (NiDB) [19], developed at Hartford Hospital for the Olin Neuropsychiatry Research Center, also provides an open-source database and a pipeline management system. It provides a platform to store and manipulate neuroimaging data and addresses challenges of data sharing identified by the International Neuroinformatics Coordinating Facility (INCF) Task Force on neuroimaging data sharing. It not only supports local storage and analysis of neuroimaging data but also provides data sharing features. NiDB also provides a search mechanism for pipeline results with predefined search criteria for which a user can provide specific values. However, it does not provide a way for the user to build a dynamic query for the clinical variables present in a dataset. The N4U Data Atlas supports this feature (discussed in Section 6). Moreover, NiDB can also support the creation and execution of pipelines on local clusters; however, it is not clear from the description whether it can also support execution of those pipelines on other distributed environments such as the Grid. In contrast, in N4U all user pipelines are executed on the Grid infrastructure and the datasets are also stored on the Grid. A user can locate the dataset's Logical File Name (lfn) based on the search criteria he or she constructed on the querying interface of the Data Atlas. This information can then be passed on to the N4U Analysis Service through CRISTAL to execute some user analysis on the selected datasets.

The Extensible Neuroimaging Archive Toolkit (XNAT) [20], developed by the Neuroinformatics Research Group at Washington University at St. Louis, aims to offer researchers an integrated environment for the archival, searching and sharing of neuroimaging datasets. It relies on an extensible XML schema to represent imaging and experimental data and supports a relational database backend. It is aimed at managing large amounts of data via a three-tier design infrastructure consisting of a client front end, the XNAT middleware and a data store. The data store comprises a relational database and a file system on which images are stored. A small portion of the schema presented in this paper has a purpose similar to XNAT i.e., to store pointers to files in the database. However, the data storage and querying mechanism of the N4U Data Atlas differ from XNAT's approach. XNAT relies upon hybrid storage i.e., XSDs and a relational database to represent and store data. XNAT generates relational databases from the XSDs, which are site dependent. This dual representation requires that changes to the data model percolate through the XSDs and the database, and this often requires manual

intervention. However, this is not the case with the N4U Data Atlas that uses a relational database only. Moreover, the primary querying language in XNAT is based on the XPath (www.w3c.org/TR/xpath) and the XQuery (www.w3c.org/TR/xquery) approaches. The N4U Data Atlas uses SQL as its query language. Furthermore, the web forms used in XNAT are also generated from XSDs schema in contrast to the dynamic form generation in the N4U Data Atlas using the values from its database.

The GridPACS system [21] supports the distributed storage, retrieval and querying of image data and associated descriptive metadata. Workflows and metadata are modeled as XML schema unlike the relational database schema approach used in the N4U Analysis Base, which supports rich querying using SQL. GridPACS integrates image data and metadata to maintain provenance information about how the images have been acquired and processed. SenseLab [6], developed at Yale University, is a metadata driven system to store scientific data using an entity-attribute-value approach with classes and relationships represented in a relational database. Most of these approaches primarily focus on the storage of clinical data. In contrast the N4U Analysis Base not only provides a schema to store and retrieve image datasets but also contains a layout for storing associated clinical study data with subjects linked to image datasets, and pipelines.

3. The N4U Analysis Base

In N4U the datasets are stored on its Grid-based infrastructure; nevertheless providing access to these datasets is not trivial because a user, who may want to carry out an analysis, may need to select part of the dataset based on some specific characteristics. Filtering gigabytes of data at runtime is a non-trivial task, unless it is methodically indexed beforehand. Furthermore, creating such indexes also becomes challenging when the metadata associated with the datasets is stored separately [22]. The challenge is further compounded by the presence of disparate formats of various datasets. The N4U Analysis Base has been designed in such a way that overcomes these challenges and provides functionality in the N4U project such as: (i) access to indexed images and datasets [23]; (ii) access to indexed workflow/pipeline and algorithm definitions; (iii) the provision of interfaces for data export into the Analysis Base; and (iv) the linkage required between neuroimages and clinical data to provide an effective data retrieval facility via *ad-hoc* queries. In this way the N4U Analysis Base offers an integrated medical data analysis environment to optimally exploit neuroscience pipelines, large image datasets and algorithms for clinical analyses. This interlinking of pipelines, algorithms, clinical variables and clinical images was one of the major aims of developing the Analysis Base to conduct neuroscience analyses and to process complex user queries. The support for user defined queries enables a user to identify the location(s) of external data stored on the Grid infrastructure. These queries involve (but are not limited to) (i) searches for datasets or image files locations, their creation dates and modification times if these files were modified over a period of time; and (ii) searches for pipeline(s) or algorithm(s) by name, or with specific search requirements (for example those pipeline(s) or algorithm(s) created by a specific user, etc.)

Some of the important constraints that appeared during the course of user discussion and which influenced the design of the Analysis Base database schema are: (a) a user can create zero to any number of pipelines (a.k.a. workflows); a pipeline can have multiple algorithms and a single algorithm can be used in multiple pipelines; and (b) a dataset, which is a collection of files, can have different combinations of image and data files. A dataset can further be organised into subcategories and assessments. The clinical data and image files will, of course, have a defined inter-relationship in the Analysis Base.

The layout of the Analysis Base schema illustrating the entities and their mutual relationships is shown as Figure 2. In this schema, the user information is linked to N4U pipelines, algorithms and datasets as stored in the Analysis Base. The *User* entity in the schema uniquely identifies a user and its

associated role that is primarily maintained to respond to queries related to user roles. Any change in user registration information is also managed. The *Group* entity defines a collection of users and helps in understanding their various possible roles such as researcher, practitioner, or administrator etc. Since each user can be associated with multiple groups and a group can have multiple users, an intermediary entity *User_Group* has been created to maintain this relationship. This *User_Group* provides the *many-to-many* relationship required between the *User* and *Groups tables*. The *Pipeline* entity maintains the definition of a pipeline such as the pipeline *version* and the pipeline *description* in the N4U community. Each pipeline is uniquely identified in the Analysis Base and each user can be associated with *one-to-many* pipelines.

An *Algorithm* in N4U (also defined in the Analysis Base schema) is basically a set of tasks that a neuroscientist wants to perform over one or more neuro-datasets. Each algorithm is assigned a unique identifier. The algorithm's physical location is external to the Analysis Base and consequently the *Algorithm* entity maintains a *Logical File Name (lfn)* to address an actual algorithm stored on the N4U Grid infrastructure. On the Grid infrastructure by the Replica Location Services (RLS) or catalogue [24] this lfn is then translated to a physical location. It may be possible that some files are not stored in the Grid infrastructure; in such cases, the *lfn* denotes a fully qualified *Uniform Resource Identifier* (*URI)* to that file. To ensure the tracking of a dataset's ownership, an identification of the owner is also stored alongside each record. Moreover, there may be cases where a physical dataset is located in a hierarchical data file system served by an HTTP server. In this instance, the root *Uniform Resource locator (URL)* of the dataset is stored. Finally, each pipeline can have multiple algorithms in it and an algorithm can be used in multiple pipelines; the intermediary entity *Pipeline_has_Algorithms* maintains this relationship (as shown in Figure 2).

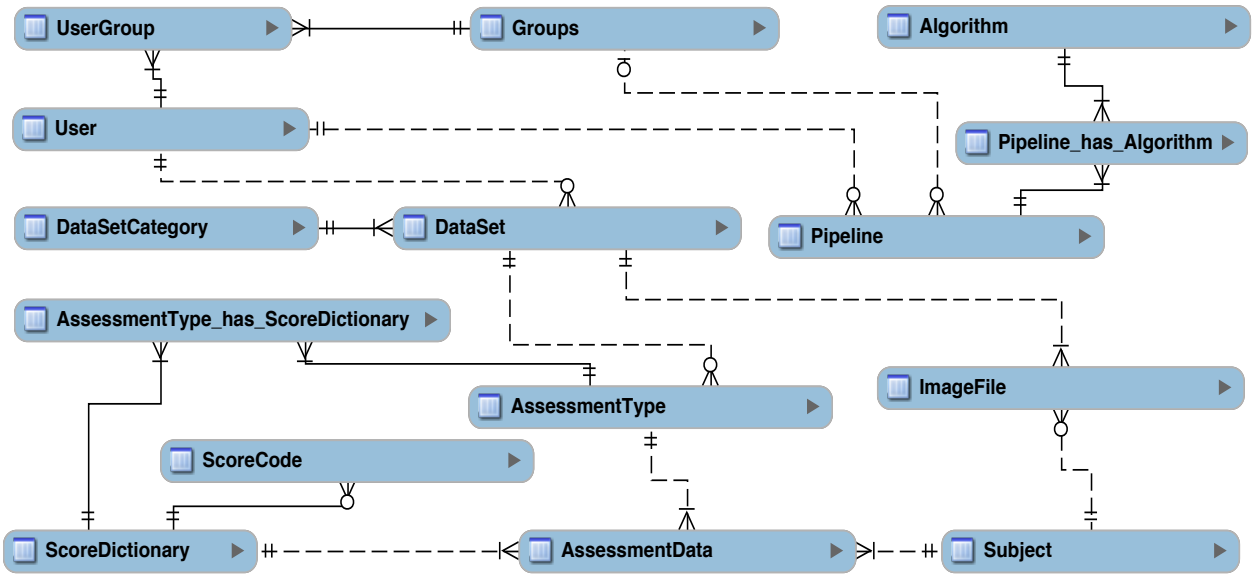


Figure 2: The Analysis Base schema diagram.

In the Analysis Base schema, the *Subject* is the entity that holds information about the patient data stored in the Analysis Base. A subject can be associated with multiple *ImageFiles,* which capture references to subjects' *brain scans* stored on the N4U infrastructure*,* each belonging to a *DataSet*. There is also a provision to bulk load image files into the Analysis Base with anonymised *Subject* information. In the Analysis Base schema the dataset definition starts with a *DataSetCategory* entity that defines each *DataSet*. For example, ADNI is a *DataSetCategory* and ADNI1, ADNI2 and ADNI-GO are various *DataSets* in the ADNI dataset category. Similarly, there can be multiple image files associated to each dataset. In

addition to image files, the Analysis Base also stores a large number of clinical variables in the *AssessmentData* that are linked to an appropriate assessment type managed by the *AssessmentType* entity. At the time of writing this paper, over two hundred thousand image files and over 10 million clinical variables data have been indexed in this Analysis Base schema.

During brainstorming workshops of requirements specifications and refinements, we came across various end-users information retrieval scenarios, where it was not obvious to the end-users how to provide values for the required clinical variables query criteria. This means that they would not know the meaning of a variable or the type of values a specific variable has in the database. For example, in one dataset there is a clinical variable called *maritalstatus*, for which the database contains numeric entries (also called scores) of 0, 1, 2, 3, 4, 5 and 9. Here, it was nearly impossible for the end-user to guess the appropriate matching number where "*Marital Status = Married*". In order to resolve the aforementioned scenario, a Clinical Variables' data dictionary has also been stored in the Analysis Base in the *ScoreDictionary* entity, which also contains clinical variables description and/or associated questions. All possible values of clinical variables (available in the *ScoreDictionary*) are stored and linked (as a *one-to-many* database relationship) in the *ScoreCode* entity. More details on the use of data dictionaries are discussed in Section 6 of this paper.

In order to populate the Analysis Base schema and to make this data available to users of N4U services, a mechanism is required that is usable by the software components which will eventually also be used by human users. In the following sections, the N4U Information Services are presented which carry out all the necessary functions of *importing*, *indexing, storing and querying* the datasets in and from the Analysis Base.

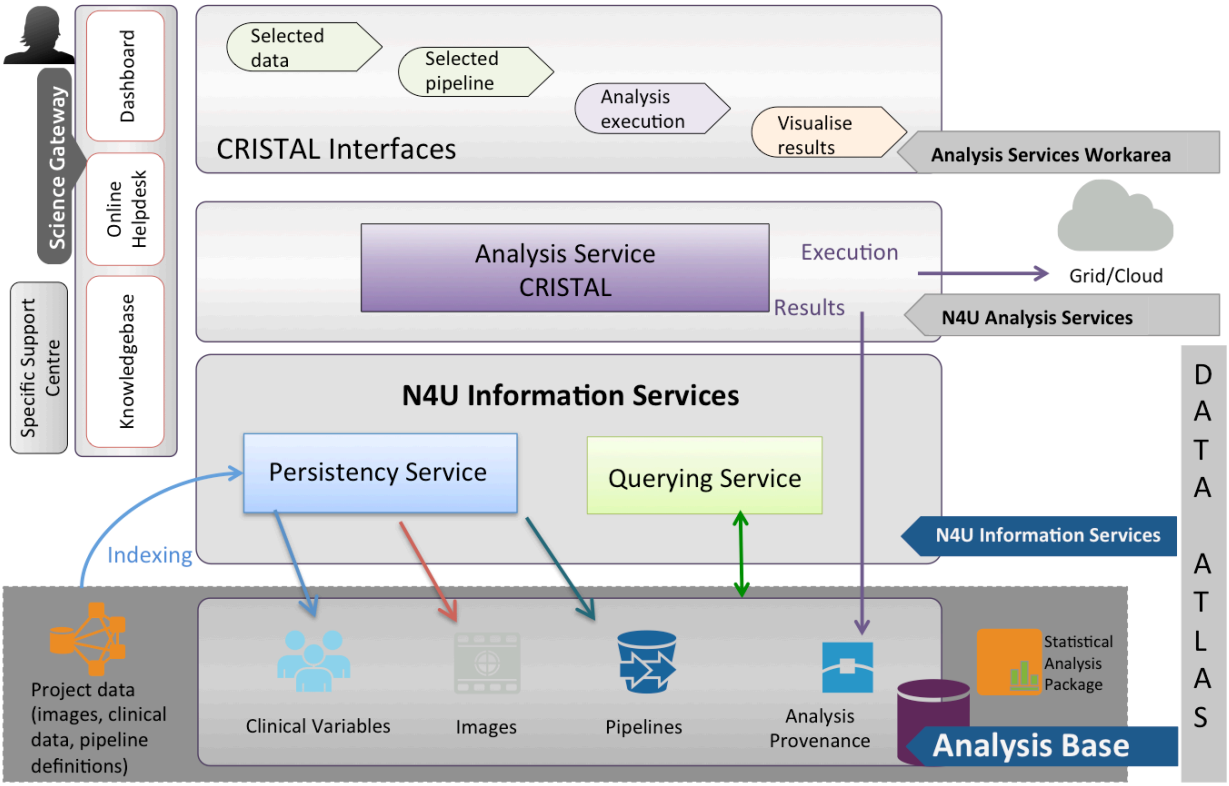


Figure 3: The N4U Virtual Laboratory architecture with essential components and services.

4. N4U Information Services

A set of integrated N4U Information Services has been defined to provide access to the underlying N4U infrastructure and to enable analyses. The Information Services first catalogue the datasets stored on the Grid into the Analysis Base and make them accessible for clinical researchers through indices. Figure 3 illustrates the detailed role of the N4U Information Services in the N4U Virtual Laboratory setup.

4.1. The N4U Analysis Base and Data Indexing Requirements and Challenges

The core data in the N4U infrastructure includes the clinical datasets from various neuroimaging studies and image files from scans carried out on patients as part of clinical studies. All datasets are imported from external sources into the N4U Grid-based storage and computing infrastructure. As noted earlier, filtering gigabytes of data at runtime is a non-trivial task, unless it is methodically indexed beforehand in the Analysis Base. The indexing of datasets is carried out to meet the following identified requirements and associated challenges:

- Neuroscientists will use the N4U Virtual Laboratory to carry out analyses by executing pre-defined scientific pipelines (also known as workflows) over different datasets.
- The N4U Virtual Laboratory needs to provide a list of the pipelines that are executable on the N4U infrastructure. The pipelines may have constraints such as being restricted to certain datasets (or formats) as inputs and algorithms that can be employed within those pipelines, etc. These constraints and other characteristics are usually specified in pipeline definitions. Therefore, the Analysis Base needs to index the pipelines registered in the N4U infrastructure. For completeness, other entities such as algorithms related to such pipelines also need to be indexed in the Analysis Base.
- In order to access the datasets stored in the N4U infrastructure, the datasets indexes need to be exported into the Analysis Base. Exporting the whole datasets into the Analysis Base is not feasible because the datasets are prohibitively large in size. For the purposes of the N4U Information Services, it is necessary to create indexes of the datasets in the Analysis Base. However, creating these indexes becomes challenging when the metadata associated with the datasets is sparse or unstructured.
- The above challenges are further compounded by the presence of disparate formats of the various datasets, without metadata in some cases, used in the N4U project. Moreover, there is no standard metadata format across the range of datasets considered in the N4U project. If such metadata existed, preferably in a standardised format that could accommodate the differences in dataset formats, it could be exported into the Analysis Base for the creation of the indices in question. Standardised interfaces for this purpose could also have been useful for accommodating any future datasets in the N4U infrastructure and their indexing in the Analysis Base.

The above-mentioned challenge of indexing heterogeneous datasets has been met by designing a generic data model for the Analysis Base (as discussed in Section 3). The Persistency Service indexes the data from the N4U Grid-based storage and computing infrastructure into the Analysis Base using this data model. The functional perspective and other important details of the Persistency Service are explained in the following sections.

4.2. The Indexing of Pipelines and Datasets via Persistency Service

The Persistency Service provides the following functionality to:

- *Crawl* the datasets in the N4U Grid storage infrastructure; or parse the user dataset sent to the Persistency Service
- *Create* a software representation of the file system storage structure that conforms to the data model of the Analysis Base;
- *Store* the clinical data (studies, results, clinical variables values/scores etc.) associated with a dataset in the Analysis Base, so that these are query-able;
- *Index* the image files within a dataset in the Analysis Base enabling association with the clinical data of that dataset;
- *Index* the pipeline definitions (and algorithms) available in the N4U infrastructure in the relevant Analysis Base storage structures; and
- *Store Metadata* associated with pipeline definitions identifying applicability of pipelines/algorithms to specific datasets.

The Persistency Service has been designed as a web service named PersistencyService deployed on the development gateway of the N4U Grid infrastructure. The PersistencyService exposes a set of operations that can be invoked by an administrator[1] or authorised N4U services[2] to execute its various functional tasks. Some of the main functional tasks of the PersistencyService are described as follows:

- **Pipelines/Algorithms Indexing**: The Persistency Service can be used to index the algorithm/pipeline definitions available in the N4U infrastructure.
- **Dataset Indexing**: To index a new dataset available in the N4U Grid storage, the administrator provides the Persistency Service with the directory structure of the dataset. Crawling through the directory structure of datasets can generate the required directory structure representation used by the Persistency Service. The service iterates through the dataset's constituent directories and builds a tree-like structure of all subdirectories and their associated contents.
- **Image Files Indexing**: In N4U, the brain images are stored in DICOM format. The Persistency Service indexes all image files encountered during the iteration of a dataset's directories. The index is made up of a fully qualified lfn that points to the location of the image file in the Grid storage.
- **Clinical Variables Data Storage**: In N4U, the clinical and subject data (e.g., patient demographics) associated with neuroimages are provided in a comma separated values (CSV) format. The CSV format was used to meet the requirements of the data provider who already had such clinical and subject data available in CSV files. The data providers upload these files onto the N4U Grid, which are stored in the dataset directories. The Persistency Service parses the contents of these CSV files and stores them in the Analysis Base along with the lfn of images.

The next section describes the process of indexing a dataset from the N4U Grid infrastructure in the Analysis Base with the help of a practical use case. At the time of writing, a number of datasets were available in the N4U Grid infrastructure that are indexed in the Analysis Base including the Open Access Series of Imaging Studies (OASIS www.oasis-brains.org) [25] (with two categories i.e. Cross-sectional and Longitudinal), the Minimal Interval Resonance Imaging in Alzheimer's Disease dataset (MIRIAD http://www.ucl.ac.uk/drc/research/miriad-scan-database), the Functional Bioinformatics Research Network dataset (FBIRN fbirnbdr.nbirn.net:8080/BDR) Phase I and Phase II [26], the European Diffusion tensor imaging study in Dementia (EDSD) [27], the Magnetic Resonance in Multiple Sclerosis dataset

[1] The PersistencyService is exposed as an interactive service to the normal users of the N4U virtual laboratory.

[2] For example, neuroscientist may create a sub-dataset to carry out an analysis via the N4U Analysis Services. Storing this user-defined sub-dataset is useful as it may speed up future analyses for the user or utilized by another user for verification of the neuroscientist's analysis. This storage may be carried out by one of the Analysis Services by invoking the relevant interfaces of the Persistency Service.

(MAGNIMS http://www.magnims.eu/), the Northwestern University Schizophrenia Data and Software Tool data (NUSDAST http://niacal.northwestern.edu/projects/9) [28], the Alzheimer's Disease Neuroimaging Initiative (ADNI http://adni.loni.usc.edu/) datasets i.e., ADNI1, ADNI2 and ADNI GO, 1000 Functional Connectomes Project (1000FCP http://fcon_1000.projects.nitrc.org/), the Alzheimer's Repository Without Borders (ARWIBO http://www.arwibo.it/), the Autism Brain Imaging Data Exchange dataset (ABIDE http://fcon_1000.projects.nitrc.org/indi/abide/), the Centre for Biomedical Research Excellence data (COBRE http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html), the Attention Deficit Hyperactivity Disorder dataset (ADHD-200 http://fcon_1000.projects.nitrc.org/indi/adhd200/) and the International Neuroimaging Data-sharing Initiative for Diffusion-weighted MRI dataset (INDI_DWI http://fcon_1000.projects.nitrc.org/indi/indi_ack.html). For the purpose of describing the functional details of the Persistency Service, we have selected the NUSDAST and FBIRN datasets as demonstrative use cases in this paper. These selected datasets also demonstrate the structural differences in the data available in the N4U infrastructure.

5. Persistency Service Dataset Indexing Use Cases

This section presents use cases of the indexing of different types of datasets in N4U. These datasets vary in their directory structure on the Grid, and also in their internal structure – i.e., how they represent the medical data along with data dictionaries and image files. It is important to understand the file and directory structure of these datasets in N4U for two key reasons. The first reason is to enable user access to the files stored on the N4U Grid infrastructure using the search mechanism provided in the Analysis Base. In order to achieve this, the Analysis Base should know about the location and path (i.e., the lfn) of these files. The second contributing reason is related to the default representation of the dataset's metadata in N4U, which is in the CSV format. In many available clinical datasets in N4U, CSV files representing the clinical variables and subject data of datasets do not maintain the image file lfns. However, they do keep track of the associated subject id. These scenarios presented a number of challenges for the Persistency Service's design and implementation and required a mechanism being devised to link clinical data from the CSV with the image files stored on the N4U Grid infrastructure.

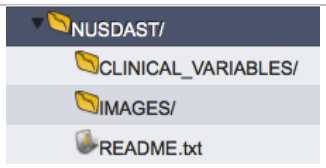
5.1. The NUSDAST Dataset

The schizophrenia community has invested substantial resources on collecting, managing and sharing large neuroimaging datasets. Its efforts have resulted in a resource known as the Northwestern University Schizophrenia Data and Software Tool (NUSDAST) that provides high-resolution magnetic resonance (MR) data from subjects with schizophrenia. The subject data also provides information about the non-psychotic siblings and their health control parameters. The NUSDAST dataset has the following two notable properties: (i) it combines neuroimaging data with demographic, clinical, neurocognitive and genotype information; and (ii) it consists of 368 subjects and for each subject there is an arbitrary number of files (minimum 3 to maximum 33) with different characteristics such as MPR, FLSH etc.

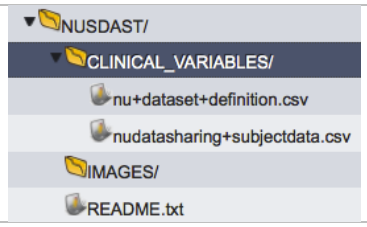
All datasets to be used in the N4U project are stored in the *data* directory of the N4U Grid infrastructure. These datasets contain both clinical study data and images. All clinical study related data (called clinical variables) are made available as variable length CSVs in the CLINICAL_VARIABLES directory. The IMAGES folder contains the image scans taken as part of the clinical studies in various formats (e.g. NIFTI .nii format). These images are organised in sub-directories that are named after the subject IDs and can thus be cross-referenced with the clinical study data of the subjects. Figure 4 illustrates the directory structure of the NUSDAST dataset in the N4U Grid storage. The .csv files in the CLINICAL_VARIABLES directory provide information about the data dictionary, which provides metadata information about the clinical variables, and the actual clinical data, which includes each subject's clinical, neuropsychological and socio-demographical variables.

Each images subdirectory is named after the scanned patient's id. The structure of scanned image files inside the subject/patient sub-directories e.g. nG+NUSDAST+CC0196, is as follows:

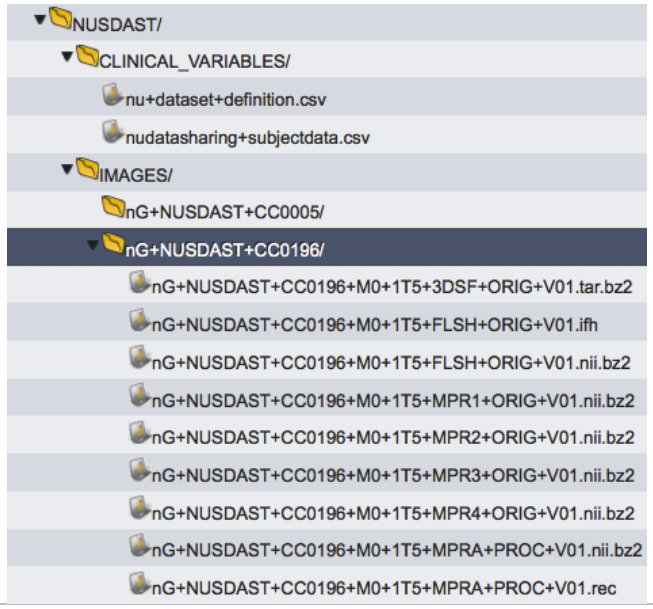
IMAGES/nG+NUSDAST+CC0196/
├── nG+NUSDAST+CC0196+M0+1T5+3DSF+ORIG+V01.tar.bz2
├── nG+NUSDAST+CC0196+M0+1T5+FLSH+ORIG+V01.ifh
├── nG+NUSDAST+CC0196+M0+1T5+FLSH+ORIG+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPR1+ORIG+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPR2+ORIG+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPR3+ORIG+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPR4+ORIG+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPRA+PROC+V01.nii.bz2
├── nG+NUSDAST+CC0196+M0+1T5+MPRA+PROC+V01.rec
…
…



a. NUSDAST top-level directory



b. NUSDAST files inside CLINICAL_ VARIABLES directory



c. NUSDAST image directories showing multiple files for a subject nG+NUSDAST+CC0196

Figure 4: NUSDAST dataset directory structure in the N4U Grid storage infrastructure.

Note that in the patient's id directory (e.g. nG+NUSDAST+CC0196 in Figure 4) the files with extensions *.rec* and *.ifh* represent the summary information of all the scans. The first four scans are MPRAGE acquired at 1.5T, while the last one is the average of the first four scans to improve the signal-to-noise ratio (SNR). The naming convention adopted for the NUSDAST scan image files is explained as follows:

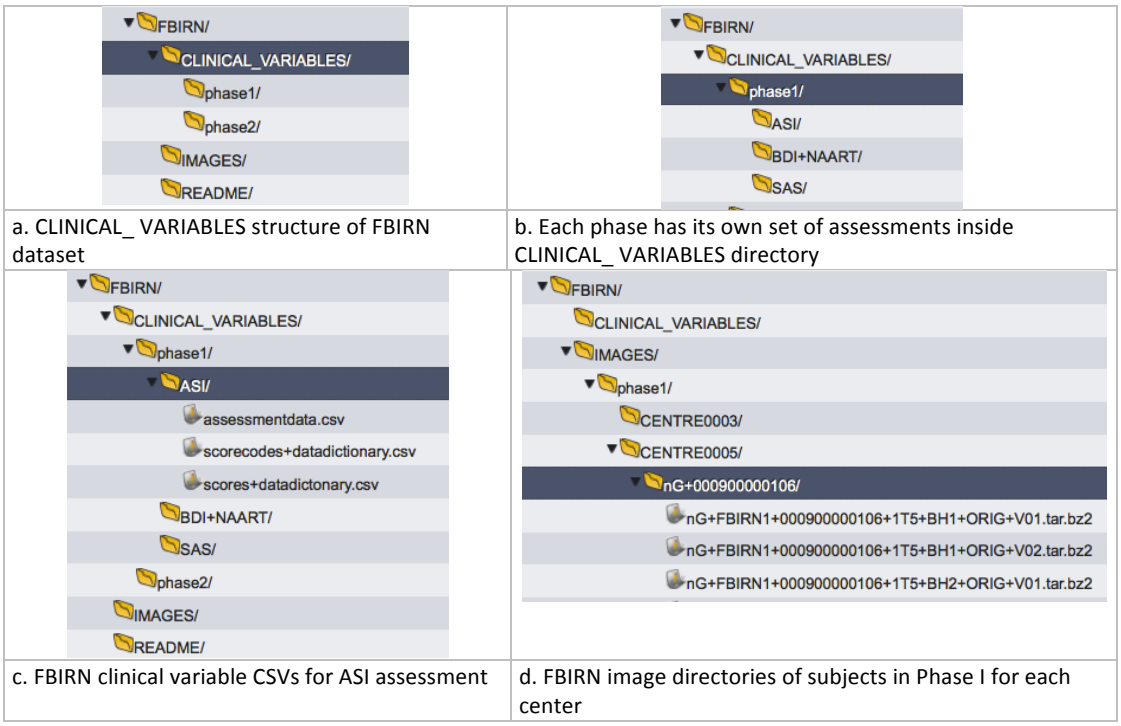
- **nG** = neuGRID
- **NUSDAST** = NUSDAST dataset
- **CC0196** = patient's id
- **M0** = Month 0 (up to 3 time points: M0, M24, M48)

- **1T5** = Field strength
- **3DSF** = 3D surface object or FLSH = FLASH or MPR1/2/3/4 = MPRAGE acquisitions or MPRA = MPRAGE average
- **ORIG/PROC** = original or processed scan
- **V01** = it indicates the version of the scan.

5.2. The FBIRN Dataset

The Biomedical Informatics Research Network [29], FBIRN dataset is hierarchical in nature, representing different phases such as Phase I and Phase II. The Phase I Traveling Subjects study was the first fBIRN multi-centre study. Five healthy subjects were imaged on two occasions on nine or ten different scanners located in geographically diverse locations. The purpose of this study was to provide a reference dataset with which to assess test-retest and between-site reliability of *fMRI* - Functional Bioinformatics Research Network, FBIRN [14]. It also provides a rich dataset to test tools and methods to allow for the calibration of between site differences in the fMRI results.

Each phase in FBIRN has its own sub-assessments as shown in the directory structure in Figure 5 (b). The .csv files representing clinical variables and metadata information are placed inside sub-directories as shown in Figure 5(c).



a. CLINICAL_ VARIABLES structure of FBIRN dataset

b. Each phase has its own set of assessments inside CLINICAL_ VARIABLES directory

c. FBIRN clinical variable CSVs for ASI assessment

d. FBIRN image directories of subjects in Phase I for each center

Figure 5: FBIRN dataset directory structure in the N4U Grid storage infrastructure.

Images in FBIRN are also arranged in a hierarchical manner representing different phases such as Phase1 and Phase2. The phase directory has further sub-directories representing the centres where the scans were performed. Different scan images of a subject are placed inside a subdirectory named after the scanned patient's identification (id). The structure of scanned image files inside the subject/patient sub-directories e.g. nG+000900000106, is as follows:

IMAGES/phase1/CENTRE0003/nG+000900000106
- nG+FBIRN1+000900000106+1T5+BH1+ORIG+V01.tar.bz2
- nG+FBIRN1+000900000106+1T5+BH1+ORIG+V02.tar.bz2
- nG+FBIRN1+000900000106+1T5+BH2+ORIG+V01.tar.bz2
- nG+FBIRN1+000900000106+1T5+BH2+ORIG+V02.tar.bz2
- nG+FBIRN1+000900000106+1T5+MMN1+ORIG+V02.tar.bz2
- nG+FBIRN1+000900000106+1T5+MMN2+ORIG+V01.tar.bz2

— nG+FBIRN1+000900000106+1T5+MMN2+ORIG+V02.tar.bz2
— nG+FBIRN1+000900000106+1T5+MN1+ORIG+V01.tar.bz2
— nG+FBIRN1+000900000106+1T5+MPR+ORIG+V01.tar.bz2
— nG+FBIRN1+000900000106+1T5+R1+ORIG+V01.tar.bz2
— nG+FBIRN1+000900000106+1T5+R1+ORIG+V02.tar.bz2
— nG+FBIRN1+000900000106+1T5+R2+ORIG+V01.tar.bz2
...
...

Note that in the patient's id directory (e.g. nG+000900000106) there are multiple files. The naming convention adopted for the FBIRN phase1 scan image files is explained as follows:

- **nG** = neuGRID
- **FBIRN1** = Dataset name
- **000900000106**= patient's id
- **1T5** = Field strength of 1.5 Tesla or 4T which means 4.0 Tesla
- **SM2** = It is the acquisition modality. Other modalities such as BH1, BH2, MMN1, MMN2, MPR, R1, R2, SIRP, SM1, SM2, SM3, SM4 and T2 are also possible.
- **ORIG** = original or processed scan
- **V01** = it indicates the version of the scan.

It is apparent from the above discussion that there are different types of datasets, which vary in structural representation on the Grid and clinical data. The FBIRN dataset is arranged in three sub-levels, the OASIS dataset is arranged in two sub-levels and the NUSDAST dataset is arranged in one sub-level only. Since there is no uniform structure across all the datasets, it is a challenge to devise a flexible mechanism that can accommodate such diversity in current datasets as well as future datasets. Based on this information and understanding, the Persistency Service devises a mechanism backed by the supporting (Analysis Base) storage design that can address this challenge. Given the above on-disk structure of the datasets in the Grid storage, their indexing in the Analysis Base by the Persistency Service is described in the next section.

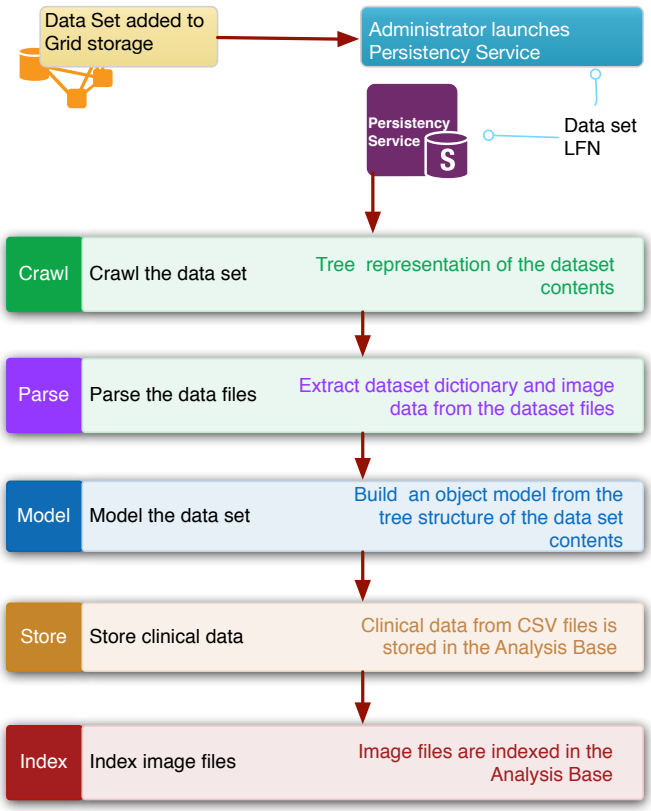


Figure 6: The Persistency Service's functional flow during the indexing of a dataset from the Grid storage.

5.3. Indexing the NUSDAST Dataset in Analysis Base

From a software design point of view, the PersistencyService consists of a crawler, parser and model components that represent the datasets according to the database schema. The crawler browses through the directory structure of a dataset and records the contents in a tree-like structure. The parser component parses the directory structure along with data dictionary files associated with a dataset. The model component transforms the clinical variable data into insert-able SQL objects by linking clinical variables with image lfns and patient ids. The SQL objects are mapped directly to the relevant tables of the Analysis Base. Figure 6 illustrates the Persistency Service's functional flow during the indexing of a dataset from the Grid storage into the Analysis Base. The details of this process are as follows.

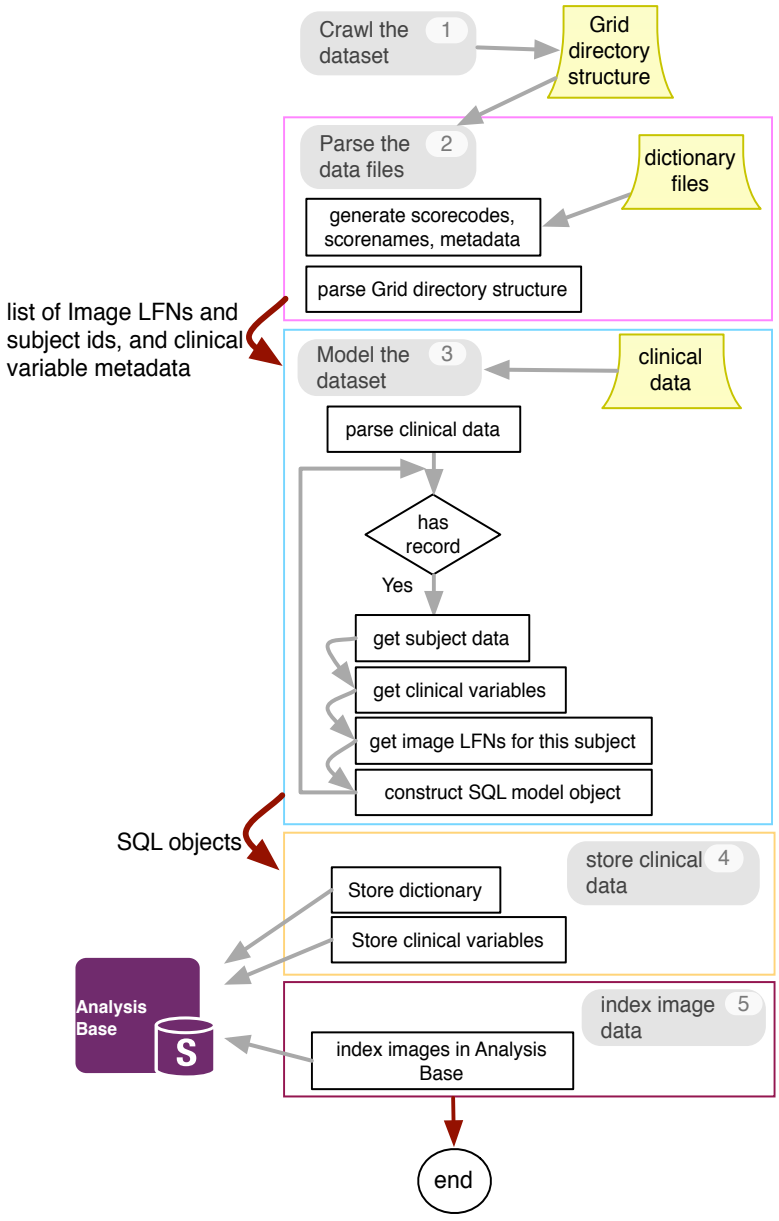


Figure 7: The Persistency Service's flow chart diagram illustrating the flow of activities during data set indexing.

The crawler creates the Grid directory structure of a given dataset. This directory structure represents the image directory and image names in each directory. This information is passed on to the parser. The parser analyses the directory structure and extracts the image lfns and the associated patient ids. It also parses the dataset dictionary, which describes the metadata of the clinical variables and possible score values (if any) for each clinical variable in the given dataset. The metadata information that is extracted from the dictionary files is stored in the Analysis Base. The image lfns, patient id (subject id) and clinical study data (in a CSV format) are passed to the model component (step 3 in Figure 6). The model iterates over the clinical study data file and retrieves a list of clinical variable values, collects subject information, and creates a mapping between a subject and its image files (passed from step 2). Once all this information is available, the PersistencyService then creates SQL models for each record according to the relevant tables that can be inserted in the Analysis Base. Figure 7 shows the Persistency Service's flow chart diagram illustrating the flow of activities during data set indexing in the Analysis Base.

The indexing of image files in the Analysis Base requires appropriate relationships to be established between each image file (its lfn to be precise) and the clinical variable's values. As explained in Sections 5.1 and 5.2, the names of the sub-directories containing the individual image files and the names of image files contain the subject IDs (i.e., the anonymised patient names). This naming convention aids in appropriately indexing the image files in the relational database model of the Analysis Base. Once the PersistencyService indexes a dataset (in terms of subjects, dataset dictionary, clinical variables and images) in the Analysis Base, other N4U services, such as the Querying Service, are able to utilise these indices in order to access the Grid-stored datasets instead of searching the full Grid-storage at runtime.

Before discussing how other services can access the datasets via these indices (Section 6), the following subsection briefly describes the indexing of algorithms and pipelines in the Analysis Base by the PersistencyService.

5.4. Persistency Service Pipeline of Algorithms Indexing Use Case

In addition to indexing datasets stored in the N4U Grid infrastructure, another function of the Persistency Service is to index pipelines of algorithms, in the form of scripts, which are also stored in the Grid infrastructure. The Persistency Service indexes the available pipelines of algorithms so that the users can make appropriate selections when defining their analyses. The following paragraph describes the mechanism used to index pipelines with their associated information.

In order to provide an exploratory and customizable search mechanism (discussed later in Section 6), the Persistency Service indexes the pipeline and its associated information. Algorithms are also conceptually contained in individual pipelines i.e., the Persistency Service indexes the algorithms while maintaining the relationship between algorithms and their encompassing pipelines. Figure 8 shows the main steps in indexing a pipeline in the Analysis Base.

After receiving the pipeline information, it is stored in a temporary file on the server. The pipeline information is then parsed to extract its name, lfn, version, description, algorithm information, etc. The parsing is performed in order to avoid null or empty fields and to build SQL objects with appropriate values. In order to store the pipeline and its associated information the SQL objects transform the given pipeline information onto the underlying schema. Once the SQL objects are created, the pipeline information is stored and indexed in the Analysis Base and then exposed to the Explorer interface, which is a part of the Querying Service (discussed in Section 6.2).

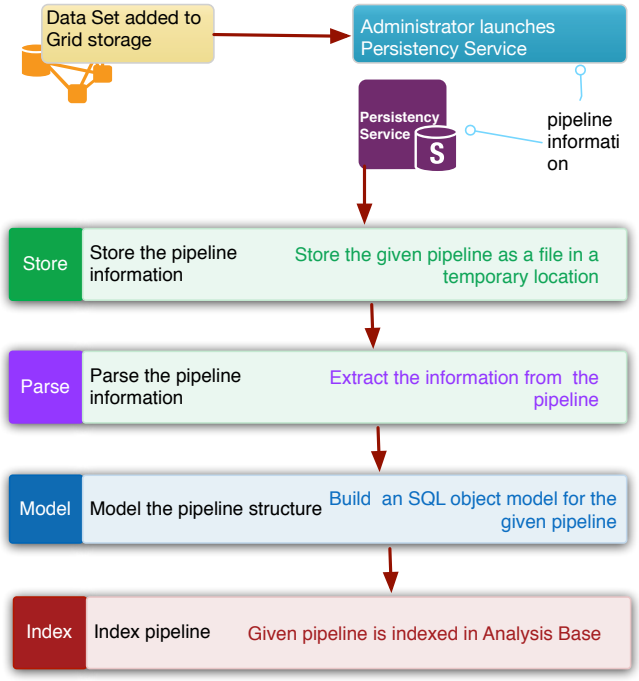


Figure 8: Main activities of the Persistency Service involved in indexing a pipeline.

6. Analysis Base Information Retrieval through the Information Services

So that N4U users and other N4U services are able to retrieve information from the Analysis Base a mechanism was required to search and query the data stored in it. This retrieval mechanism enabled users to perform their queries on clinical data and images, and to retrieve desired data related to pipelines of algorithms. Various web interfaces, called the Analysis Base Querying Interfaces, have been deployed to fulfil these requirements. Through these interfaces, the Querying Service supports a wide range of queries including browsing datasets, viewing data dictionary, searching for metadata in the patients' clinical data, searching for particular neuroimages with specific characteristics, etc. These queries may arise from different origins such as from the set of N4U services (as shown in Figure 9) and thus the Querying Service exposes different interfaces for different clients. A summary of different functions offered by the various interfaces of Querying Service is listed below:

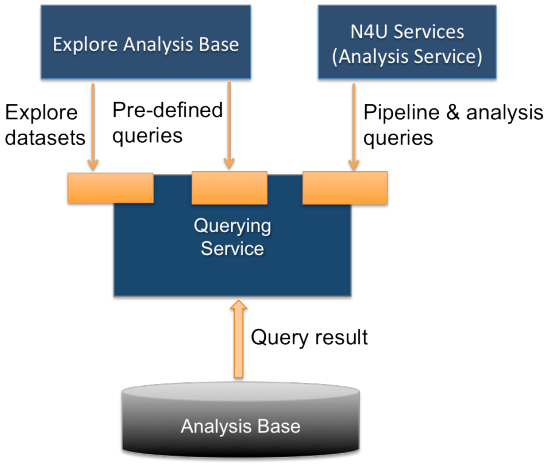


Figure 9: Users and other services accessing information using the Querying Service interfaces.

- Using the Querying Service, users can search within multiple datasets based on different search criteria and use the resultant data in their analyses.
- A few examples of supported datasets queries include: (i) searching for all or specific datasets; (ii) getting a dataset identified by an unique id; (iii) searching in a dataset based on the given values of various clinical variables; and (iv) retrieving the lfn location of an image, etc.
- A user can search for and catalogue a particular subset of image and data files residing in the N4U Grid infrastructure. This function allows for an easy-to-use mechanism for the users to access the information in the Analysis Base from a single point of access.
- Performing exact match and SQL *'like'* operations for filtering the clinical variable data. Users can also perform comparison operations (i.e., >, <, = etc.) on the values of clinical variables using the visual Query Builder.
- A user can navigate a dataset (or subsets) and view all the clinical variables linked to that dataset. A user can also view the metadata and/or detailed data dictionary associated with each clinical variable.
- The ability to export the output of user queries/inquiries about clinical data and images as both .xml and .csv files allows subsequent processing of data in other applications (e.g. Excel and Analysis Service) and also importation into user-owned databases.

In the following subsections we discuss these points in more detail along with the design and implementation aspects of the Querying Service's sub-components.

6.1. Querying Service Design and Implementation

The Querying Service is designed as a web service to achieve its multiple objectives. A web service, invoked remotely over HTTPS, allows us to expose the desired functionality in a controlled manner for the client applications. All the functionality embedded at the server-side not only provides transparent access to the service functionality but also necessitates only a small footprint at the client side. This setup allows for a rich and easy-to-use but controlled functionality for users and external services. The N4U user or other services interact with the Analysis Base through a service-oriented approach that is illustrated in Figure 10.

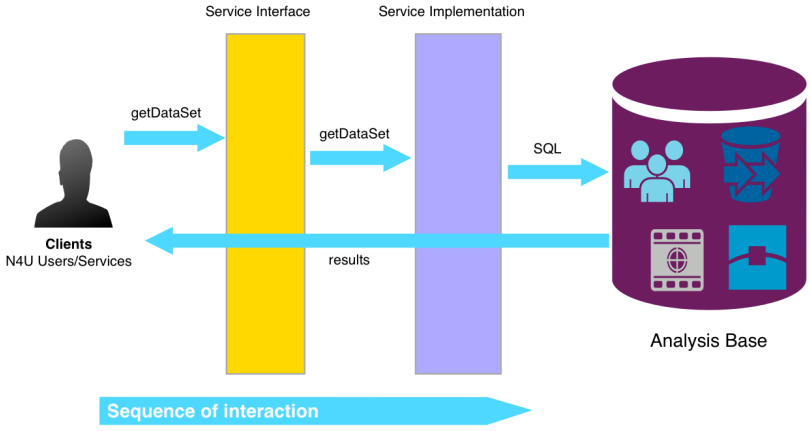


Figure 10: The user interaction with the Analysis Base through a service-oriented approach.

The Querying Service has been developed using the Apache CXF framework (http://cxf.apache.org/). The Querying Service is only accessible over HTTPS on the N4U gateway and only valid N4U users with active, authenticated sessions are allowed to access this service. Its implementation makes use of a user identity retrieval mechanism provided by the N4U gateway's runtime environment (https://neugrid4you.eu/web/science-gateway), which authenticates all incoming requests. This

approach provides two advantages: (i) non-legitimate requests cannot access the N4U data, thus making it secure and (ii) the user identity is used to log and retrieve information stored by that user in the Analysis Base. This request-response flow is illustrated in Figure 11.

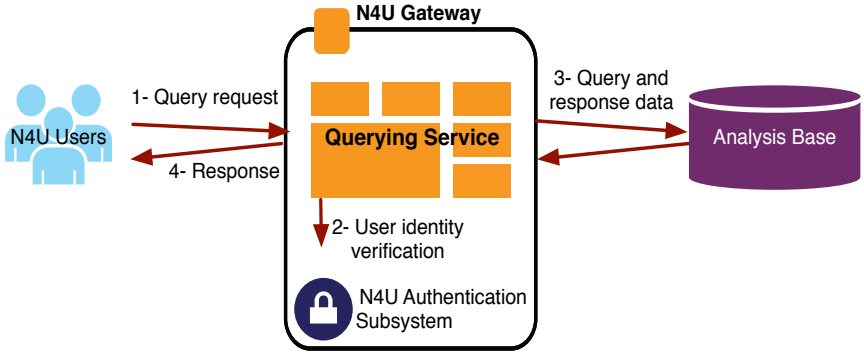


Figure 11: Flow of activities in the Querying Service.

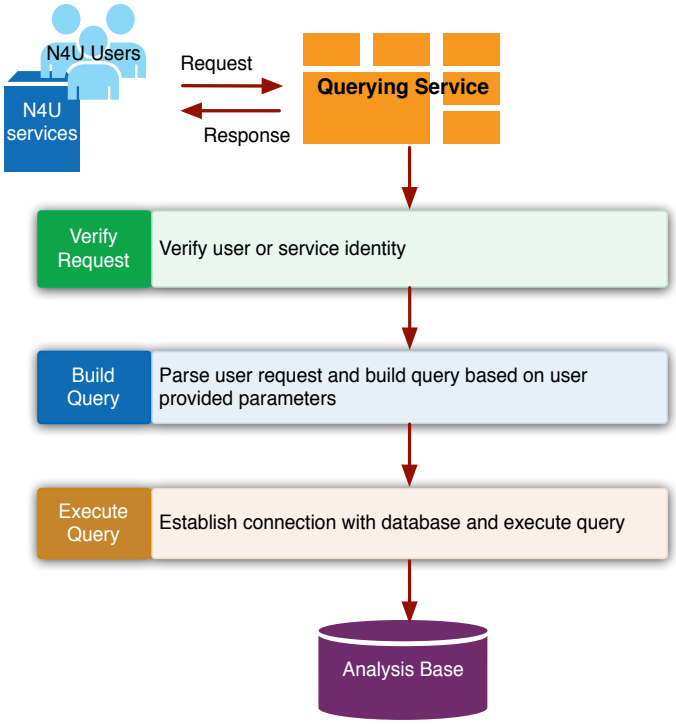


Figure 12: The Querying Service's functional flow in handling a query request.

After the successful verification of each user request, the Querying Service starts building the appropriate search query against the user parameters provided in the request. In the query-building phase, user selected parameters are parsed, and the required tables and their appropriate joins are selected based on the given parameters. The Querying Service then connects to the backend database and executes the query. The query result is retrieved from the database and presented to the user. In the case of failures e.g. user verification or query errors, human readable and understandable messages are sent back to the user. The internal flow of activities within the Querying Service is shown in Figure 12.

In order to test the Data Atlas functionality using the Querying Service interfaces for various types of users, a dedicated work package has been designed in N4U to specify user requirements and evaluation framework – User Acceptance Testing (UAT) - that describes user requirements and verifies the developed services. The UAT has been performed in 43 sessions in different locations such as Stockholm, Brescia, Amsterdam, Kuopio, Munich, Warsaw and Bern in Europe with the help of different types of N4U user group such as Neuroscientists, Pharmascientists, Developers and Administrators. These users are divided into those internal and external to the N4U infrastructure categories. The internal users are the ones who are participants or developers of the N4U services, and the external users are predominantly the neuroscience and pharmascience researchers. The details about user requirements, their evaluation framework, training and user acceptance testing are presented in the relevant project deliverables [30]. The results of the UAT confirmed that all of the essential requirements such as the interface for basic searches and the interface for advance searches related to the Querying Service and Data Atlas have been successfully achieved. The following sections discuss different Analysis Base querying interfaces provided by the Querying Service.

6.2. Information Retrieval via Analysis Base Explorer

Whilst designing various Analysis Base querying interfaces, special focus was placed on increasing the ease of learning for novice users as well as providing intelligent dataset filtering features for expert and advanced users. The main purpose of having a querying mechanism on top of the Analysis Base was to facilitate a user or external services in fetching the data stored in it. However, not every user – particularly a neuroscientist - is equipped with database querying skills and relevant technical expertise. Because of this, a feature rich web interface has been designed and developed to present the N4U Analysis Base to the N4U community. The *Analysis Base Explorer* presents an accessible and easily usable view of the N4U Analysis Base from the perspective of a naive user. It is designed to showcase the main datasets, sub-datasets and pipelines stored in the Analysis Base. When a user accesses the Explorer page, two main entities i.e. the datasets and pipelines are presented. Upon selecting one of these entities, an AJAX (Asynchronous Java and XML) request is sent to the Querying Service to retrieve further information (data) about the selected entity. The resultant information is displayed which also includes the total number of records as well as the query response time, and this further improves the users experience. Through this Explorer interface, users can also observe the relationship between pipelines and algorithms. For instance, when a user selects Algorithms after selecting a Pipeline, the data representing their relationship are also displayed. This kind of information is particularly useful for neuroscientists looking for a specific algorithm in a particular pipeline.

Data stored in the Analysis Base can massively grow over time and the links between different entities can also cause a large amount of data to be retrieved, and this presented a challenge in retrieving and presenting the data at the client side. To safeguard against this potential problem, a pagination technique has been adopted through which the data is divided into a configured number of records (set to 300 records per page by default) when presented to the user. This not only decreases the burden on the backend database server answering user queries but also significantly reduces the response time and rendering time at the client/user side.

6.3. Information Retrieval via Analysis Base Querying Builder

In the previous section, we have discussed how the *Explorer* component of the Querying Service assists novice users in viewing the data stored in the N4U Analysis Base. A user, such as a neuroscientist, may also want to retrieve the neuroimaging data, filtered on the basis of certain clinical variables. To support this type of complex search and selection, a dynamic and user driven *Query Builder* has been incorporated into the web interface. The Query Builder allows a user to graphically design queries based on his or her selection criteria to search for the images within the datasets indexed in the Analysis Base.

In order to further facilitate N4U users, three different types of graphical query building methods are provided: (a) Predefined queries; (b) Clinical variable searches; and (c) Advanced searches (as shown in Figure 13). These methods are explained in following subsections.

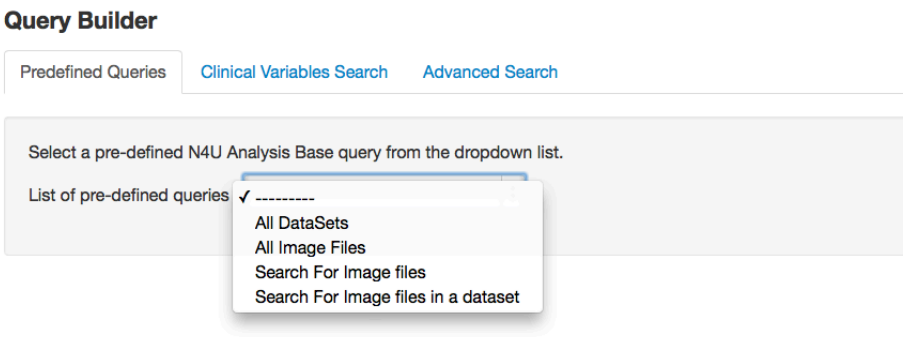


Figure 13: Query Builder interface to design queries and retrieve data from the Analysis Base.

6.3.1. Predefined Queries

In the Predefined Queries tab, a pre-populated list of predefined queries is provided to the user to fulfil basic user queries. The reason for having such a list of predefined queries is to accommodate a number of common use cases and to save users' time in creating their own queries for common scenarios. A dropdown list of queries is given to the users from which they can select an appropriate query. A brief description of these queries is given below.

- *All DataSets:* This query fetches a list of datasets and their associated metadata. The metadata contains dataset name, id, *lfn*, owner, creation date etc.
- *All ImageFiles:* This query retrieves a list of all image files indexed in the N4U Analysis Base. Since there could be a large number of image files, the result is divided into pages for performance reasons as discussed earlier in Section 5.2.
- *Search for Image files:* In this option, a text field is presented to the user in which he or she can provide an image filename or part of an image's *lfn*. This query searches for the given input in image names or *lfns* in all datasets and retrieves a list of matched image files.
- *Search for Image files in a dataset:* In this option, the user provides the image name or its *lfn* along with the dataset type in which the user wants to look for a given image name or *lfn*. The Querying Service looks for the given image file in the provided dataset only.

The list of predefined queries is easily extendable and administrators can add further queries for common scenarios.

6.3.2. Clinical Variable based Search

For the clinical variable based search, the Query Builder records various input parameters in a stepwise manner. Firstly, it shows a list of available datasets to the user. Each of the datasets can have multiple subcategories or sub-datasets. Upon selecting one of the datasets, the Query Builder dynamically loads the associated sub-datasets e.g. the OASIS-CrossSection and OASIS-Longitudinal for the OASIS dataset. Each subtype can have a different number of clinical variables, possibly with different names. Upon subtype selection, its associated clinical variables are dynamically created and presented as a dropdown list. In N4U, it is also possible that a dataset such as NUSDAST does not have further subcategories. For such a dataset, clinical variables associated with it are loaded and presented in a dropdown list to a user. Neuroscientists can then select from the available clinical variables and provide the selection criteria by filling the text fields. The interface allows a user to specify a multi-parameter

query by selecting and providing search criteria for multiple clinical parameters. The user may also remove certain parameters while the query building is in process or restart from scratch. As shown in Figure 14, all of the associated clinical variables for the user selected dataset 'NUSDAST' are shown to the user in order to perform a selection of multiple clinical variables and to create an appropriate search filter.

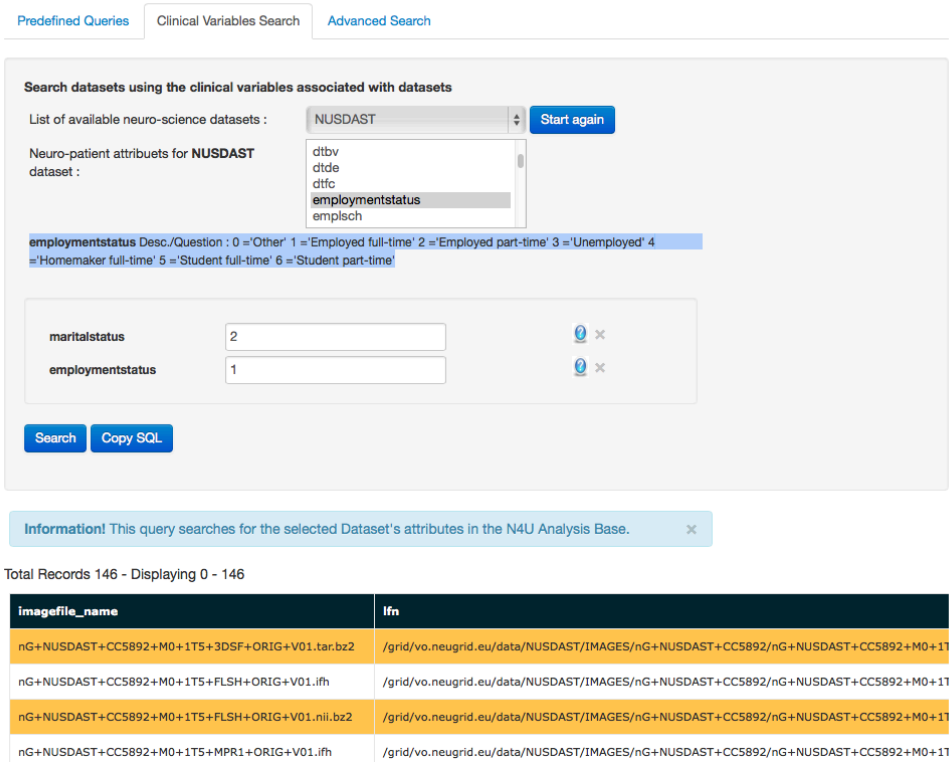


Figure 14: Clinical variables-based parameterized querying and the use of the Data Dictionary.

6.3.3. Use of Metadata and Data Dictionary for Clinical Variables based Search

Whilst implementing and testing the Clinical Variables search, we came across various scenarios where it was not obvious for the end-users how to provide values for the selected clinical variables such as *maritalstatus*. This means that they would not subsequently know the meaning of the clinical variable or the type of values that the selected variable has in the database (as discussed earlier in Section 3). For example in NUSDAST and ADNI datasets there is a clinical variable called *maritalstatus* i.e., "Marital Status", for which the database contains numeric entries as 0, 1, 2, 3, 4, 5 and 9. Here, it was nearly impossible for the end user to guess the appropriate matching number where "Marital Status = Married".

In order to resolve the aforementioned scenario, Clinical Variable data dictionaries have also been stored in the Analysis Base and their relationship has been established with each Clinical Variable for all datasets. Due to this feature, when a user selects a Clinical Variable such as "Marital Status", the respective data dictionary values are automatically shown to the user as follows (see Figure 14):

Clinical Variable: maritalstatus
Desc./Question:
0 ='Other'
1 ='Single'
2 ='Married/common law'
3 ='Divorced'
4 ='Separated'
5 ='Widowed'

9 ='Unknown'

Consequently, if the user is interested in only retrieving those subjects with *Marital Status='Married'*, then he or she can establish this from the data dictionary entries and enter 2 in the text field (as shown in Figure 14). Similarly, for another similar Clinical Variable "*employmentstatus*" (i.e. Employment Status), the following are the associated data dictionary entries:

Clinical Variable: employmentstatus
Desc./Question:
0 ='Other'
1 ='Employed full-time'
2 ='Employed part-time'
3 ='Unemployed'
4 ='Homemaker full-time'
5 ='Student full-time'
6 ='Student part-time'

Sometimes there is additional linked metadata information to a clinical variable e.g. *Variable Type, Score Values and/or Comments*. In order to view such detailed information associated with a clinical variable, the user can click on the "?" button (as shown in Figure 14), which is provided on the right-hand side of each selected variable, to display the complete metadata stored in the Analysis Base.

6.3.4. Advanced Search for Expert users

While building queries/filters using the Query Builder interface, a user can use various types of operators such as = (*equal to*), < (*less than*), > (*greater than*), *Like* (for sub-string matching), " " (for exact string match), etc. Here the *equal* operator '=' compares single values to one another in a SQL statement. The *equal* sign (=) symbolizes equality. When testing for equality, the compared values must match exactly or no data is returned. Similarly the < (*less-than*) and > (*greater-than*) are used with numerical values, and the *Like* operator can be used to treat the given text input as a sub-string to be matched anywhere in the clinical variable value. Within the Analysis Base web application detailed instructions have been provided on how to use these operators; users can also consult the public training videos provided as part of the learning material [31] to use such functions.

In addition, a user can easily specify more complex criteria (e.g. use of *OR* and *NOT Equal To, NOT IN* operations etc.) by copying the previously generated SQL and using it in the Advance Search tab. This feature is for advanced users who have an understanding of databases and know how to specify a SQL query. A text area is provided where they can specify database queries to fulfil their requirements. In order to facilitate such users, a mechanism is provided which negates the requirement for queries to be constantly rewritten; rather they can copy their previously formulated query via *Query Builder* and modify it within the *Advance Search tab*. In this way, it reduces the chances of an error, and takes less time and effort on the user's part to edit and/or extend a search criterion. Moreover, a user can only change the *WHERE* clause to modify the given filter or introduce further filtering (using OR, NOT etc. operators) as well as including further clinical variables in the overall search criteria. To avoid SQL injection [32] and other write-able queries such as INSERT, UPDATE or DELETE, the Querying Service executes these queries within a sandbox which is created by parsing the given query and using the read-only permissions on the database.

6.3.5. Exporting query results in XML and CSV format

In the above sections, we have listed a subset of the mechanisms for users to locate their desired datasets in the Analysis Base. Another important feature that this section describes is the exporting of search results or resultant outcomes in XML or CSV formats, which can then be used by neuroscientists in any preferred neuroscience data analysis application(s).

Using the Query Builder interfaces, once the user has entered search criteria and located the desired datasets, in addition to displaying the querying result, the user is also provided with the options of exporting the outcome in both CSV and XML formats. Users can do this without navigating away from the Query Builder interface and can generate a number of exports by just changing the filtering criteria. In order to enable fast data retrieval and the generation of XML and CSV, secondary B+ trees and bitmap indexing have been implemented on the data stored in the Analysis Base. When a user submits a search query based on some specified criteria, the retrieved outcome is displayed and both the CSV Export and XML Export buttons appear automatically at the bottom of the page. Upon clicking the export buttons, an export request is sent to the server. The server processes the request, generates and transforms the resultant data in the requested format and returns it as a file (.csv or .xml). Figure 15 shows the generated and exported XML file. Each generated file is time stamped, which is unique for all files; therefore the old generated files are not automatically replaced with new files. The generated XML structure provides information about an image file and its metadata, dataset and assessment, and its associated clinical variables, which were specified in the query. The following example shows one XML record from the exported XML file in which individual records are separated with the *<Record> and <\Record>* tags, and data for the clinical variables are enclosed within individual named tags.

```
<Record>
    <imagefile_name>nG+NUSDAST+CC7959+M0+1T5+3DSF+ORIG+V01.tar.bz2</imagefile_name>
        <lfn>/grid/vo.neugrid.eu/data/NUSDAST/IMAGES/nG+NUSDAST+CC7959/nG+NUSDAST+CC7959+M0+1T5+3DSF+ORI
        G+V01.tar.bz2</lfn>
        <imagefile_type>.bz2</imagefile_type>
        <imagefile_description></imagefile_description>
        <added_on></added_on>
        <dataset_id>32</dataset_id>
        <subject_id>nG+NUSDAST+CC7959</subject_id>
        <assessment_type>NUSDAST</assessment_type>
        <maritalstatus>4</maritalstatus>
        <race>2</race>
        <gender>male</gender>
    </Record>
```

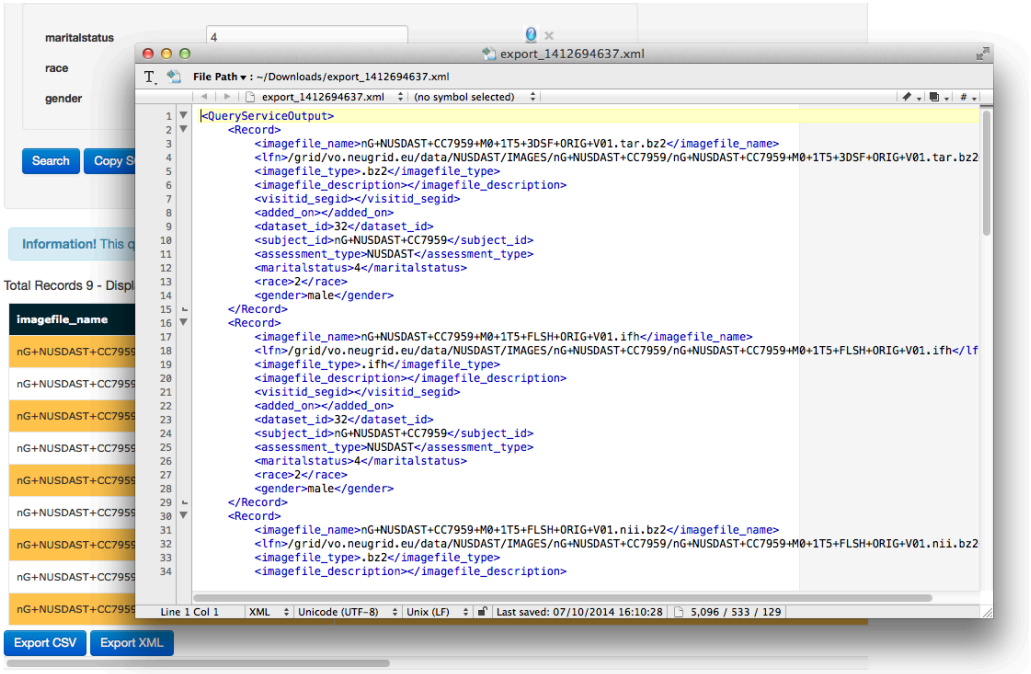


Figure 15: An example of XML export for the filtered datasets.

Whenever a user exports the query result in an XML format, seven default values appear in all the records. These seven values are *imagefile_name, lfn, imagefile_type, imagefile_description, added_on, dataset_id, subject_id,* and *assessment_type*. Here *imagefile_name, lfn*, and *subject_id* are the most important variables in order to locate a particular image on the N4U Grid infrastructure. The rest of the clinical variables, which become part of the exported XML, depend on the user provided query criteria. As discussed above, a user can include unlimited numbers of clinical variables as search filters and accordingly all of them appear in the exported file. Finally, the outcome exported in the format of XML and CSV can be used by other applications to generate/perform an analysis or to retrieve an image or set of images from the Grid.

7. Conclusions

Recent developments in data management and imaging technologies have significantly affected diagnostic and extrapolative medical research and biomedical researchers are faced with severe problems of heterogeneous clinical data management. The impact of these new technologies is largely dependent on the speed and reliability with which the medical data can be visualised, analysed and interpreted. Grid computing, and recently Cloud computing, has alleviated some of the issues associated with the capture, processing and storage of huge numbers of medical images and associated clinical information. However, there is still a lack of clinician-friendly graphical interfaces and tools that provide an easy access to the (Grid) infrastructure-resident data. Moreover, there are no or few links between images, data dictionaries, clinical data and metadata associated with images, which makes it extremely difficult for biomedical researchers to define and conduct their analyses.

In order to address these issues, the N4U Virtual Laboratory presented in this paper focuses on providing an *intuitive, fast and linked* access to large-scale heterogeneous clinical datasets to derive a greater understanding of the neuro-degenerative diseases through data analysis. The N4U Data Atlas within the N4U Virtual Laboratory addressed the research and practical challenges (discussed in Section 4). In doing so, it offers an integrated medical data analysis environment to optimally exploit neuroscience pipelines containing algorithms, large image datasets and clinical variables data to facilitate analyses. The Analysis Base enabled such analyses by indexing and interlinking the neuroimaging and clinical study datasets, and pipeline definitions stored in the N4U Grid infrastructure. Furthermore, this paper has described the requirements, specification, implementation and deployment of the N4U Information Services i.e. (1) the indexing of pipelines/algorithms and heterogeneous datasets in the Analysis Base by the Persistency Service, with a particular emphasis on how the datasets that are stored in the N4U Grid-based storage infrastructure are indexed homogeneously in the Analysis Base; and (2) the retrieval of indexed information from the Analysis Base through the Querying Service; for which various (including dynamic) interfaces and methods are exposed to provide Analysis Base access to the end-users and other N4U services.

The main challenges tackled in providing information services relate to the very large datasets sizes, different formats, structures and semantics of datasets. Moreover, the manner in which relationships between clinical variables in CSV files and image files are maintained also varies across datasets. Additionally, unstructured data dictionaries compound the problem. To overcome these challenges, a generic schema model of the N4U Analysis Base has been presented; and a Persistency Service has been presented to index and link such data in the Analysis Base. Whilst doing so, it has been specifically ensured that the users, and other N4U services, are provided with a uniform view of the datasets and data dictionaries in order to apply filters on available clinical parameters and to retrieve relevant information.

In the future, this work can be extended to resolve various challenges associated with the management of large clinical datasets. As the volume and variety of datasets in terms of different

formats and structures, can increase in future the domain becomes a Big Data candidate. Consequently, the use of *NoSQL* databases to provide fast and scalable storage and retrieval mechanisms along with biomedical Big Data analytics and information visualization can be explored. Likewise, the use of ontologies with domain knowledge on top of the Analysis Base can be explored to provide semantics which will help in achieving domain-based or keyword-based search performance, the creation of reference data and to enable reasoning. One of the best possible ways to achieve this is by building an ontological knowledge base of large-scale (Big) datasets, which includes the definition of a semantic model, the specification of domain knowledge, and the definition of links between different types of semantic knowledge.

Acknowledgments

The authors would like to acknowledge the support of the European Union in funding this work via the neuGRID4You (N4U) project (grant agreement n. 283562, 2011-2014), with special thanks to the N4U consortium.

References

[1] The neuGRID4You (N4U) project. (2015). Available from: https://neugrid4you.eu/ [Accessed 10-July-2015].

[2] CRISTAL software. (2015). Available from: http://cristal-ise.org [Accessed 10-July-2015].

[3] Spulber, G., Damangir, S., and Wahlund, L. O. (2012). D3.1: N4U requirements for the provision of knowledge management services and analysis environment, neuGRID consortium.

[4] Heinis, T., Branco, M., Alagiannis, I., Borovica, R., Tauheed, F., and Ailamaki, A. (2011). Challenges and opportunities in self-managing scientific databases. IEEE Data Eng. Bulletin, 34(4): 44–52.

[5] Cheung, K. H., Lim, E., Samwald, M., Chen, H., Marenco, L., Holford, M. E., Morse, T. M., Mutalik, P., Shepherd, G. M., and Miller, P. L. (2009). Approaches to neuroscience data integration. Briefings in Bioinformatics, 10(4): 345–353.

[6] SenseLab. (2015). Available from: https://senselab.med.yale.edu/ [Accessed 10-July-2015].

[7] Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Muller, H. M., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C., and Williams, R. W. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. Neuroinformatics, 6(3): 149–160.

[8] Marenco, L., Wang, R., Shepherd, G. M., and Miller, P. L. (2010). The NIF DISCO framework: Facilitating automated integration of neuroscience content on the web. Neuroinformatics, 8(2): 101–112.

[9] Gupta, A., Ludascher, B., and Martone, M. (2000). Knowledge-based integration of neuroscience data sources. In Scientific and Statistical Database Management, 2000. Proceedings. 12th International Conference on, pages 39–52.

[10] Samwald, M., Lim, E., Masiar, P., Marenco, L., Chen, H., Morse, T., Mu- talik, P., Shepherd, G., Miller, P., and Cheung, K.H. (2009). Entrez neuron rdfa: A pragmatic semantic web application for data integration in neuroscience research. Studies in health technology and informatics, 150:317.

[11] Fong, L., Larson, S. D., Gupta, A., Condit, C., Bug, W. J., Chen, L., West, R., Lamont, S., Terada, M., and Martone, M. E. (2007). An ontology-driven knowledge environment for subcellular neuroanatomy. In Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions, Innsbruck, Austria, June 6-7, 2007.

[12] Michel, F., Gaignard, A., Ahmad, F., Barillot, C., Batrancourt, B., Do- jat, M., Gibaud, B., Girard, P., Godard, D., Kassel, G., Lingrand, D., Malandain, G., Montagnat, J., Pelegrini-Issac, M., Pennec, X., Ro- jas Balderrama, J.,

and Wali, B. (2010). Grid-wide neuroimaging data federation in the context of the neurolog project. Stud Health Technol Inform, 159: 112–123.

[13] McClatchey, R., Branson, A., Anjum, A., Bloodsworth, P., Habib, I., Munir, K., Shamdasani, J. and Soomro, K. (2013). Providing traceability for neuroimaging analyses. International Journal of Medical Informatics, 82(9): 882–894.

[14] Ozyurt, I., Keator, D., Wei, D., Fennema-Notestine, C., Pease, K., Bockholt, J., and Grethe, J. (2010). Federated web-accessible clinical data management within an extensible neuroimaging database. Neuroinformatics, 8(4): 231–249.

[15] Small, S. L., Wilde, M., Kenny, S., Andric, M., and Hasson, U. (2009). Database-managed grid-enabled analysis of neuroimaging data: the cnari framework. International Journal of Psychophysiology, 73(1): 62–72.

[16] Hasson, U., Skipper, J. I., Wilde, M. J., Nusbaum, H. C., and Small, S. L. (2008). Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. Neuroimage, 39(2): 693–706.

[17] Zhao, Y., Hategan, M., Clifford, B., Foster, I., von Laszewski, G., Nefedova, V., Raicu, I., Stef-Praun, T. and Wilde, M. (2007). Swift: fast, reliable, loosely coupled parallel computation, IEEE Congress on Services, 9-13 July 2007, Hawaii, USA, pp. 199-206.

[18] Keator, D.B., Ozyurt, B., Wei, D., Gadde, S., Potkin, S.G., Brown, G., Morphometry BIRN, FBIRN., Grethe, J. (2006). A General and Extensible Multi-Site Database and XML based Informatics System for the Storage, Retrieval, Transport and Maintenance of Human Brain Imaging and Clinical Data. Annual Meeting of the Organization for Human Brain Mapping, Florence, Italy.

[19] Book, G. A., Anderson, B. M., Stevens, M. C., Glahn, D. C., Assaf, M., and Pearlson, G. D. (2013). Neuroinformatics database (NiDB) – a modular, portable database for the storage, analysis, and sharing of neuroimaging data. Neuroinformatics, 11(4): 495–505.

[20] Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics, 5(1): 11–34.

[21] Hastings, S., Oster, S., Langella, S., Kurc, T. M., Pan, T., Catalyurek, U. V., and Saltz, J. H. (2005). A grid-based image archival and analysis system. Journal of the American Medical Informatics Association: JAMIA, 12(3): 286–295.

[22] Munir, K., Kiani, S. L., Hasham, K., McClatchey, R., Branson, A., Shamdasani, J. and the N4U Consortium, (2013). An Integrated e-Science Analysis Base for Computation Neuroscience Experiments and Analysis, In Elsevier's journal of Procedia - Social and Behavioral Sciences, Presented at International Conference on Integrated Information, IC-ININFO 2012, Aug 30 - Sep 3, 2012, Volume 73, 27, Pages 85-92, ISSN 1877-0428.

[23] Munir, K., Kiani, S. L., Hasham, K., McClatchey, R., Branson, A., Shamdasani, J. (2014). Provision of an integrated data analysis platform for computational neuroscience experiments, Journal of Systems and Information Technology, Vol. 16 Issue: 3, pp.150 – 169.

[24] Chervenak, A., Palavalli, N., Bharathi, S., Kesselman, C., and Schwartzkopf, R. (2004). Performance and scalability of a replica location service. In proceedings of 13th IEEE International Symposium on High performance Distributed Computing, pages 182–191.

[25] Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open access series of imaging studies (oasis): Longitudinal MRI data in nondemented and demented older adults. Journal of cognitive neuroscience, 22(12): 2677–2684.

[26] Glover, G. H., Mueller, B. A., Turner, J. A., van Erp, T. G. M., Liu, T. T., Greve, D. N., Voyvodic, J. T., Rasmussen, J., Brown, G. G., Keator, D. B., Calhoun, V. D., Lee, H. J., Ford, J. M., Mathalon, D. H., Diaz, M., O'Leary, D. S., Gadde, S., Preda, A., Lim, K. O., Wible, C. G., Stern, H. S., Belger, A., McCarthy, G., Ozyurt, B., and Potkin, S. G. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional mri studies. J Magn Reson Imaging, 36(1): 39–54.

[27] Teipel, S., Dyrba, M., Frisoni, G., Bokde, A. L., Fellgiebel, A., and Hampel, H. (2012). The European diffusion tensor imaging study in dementia (edsd): Physical phantom study and multimodal clinical study on the diagnosis of alzheimer's disease.

[28] Wang, L., Kogan, A., Cobia, D., Alpert, K., Kolasny, A., Miller, M. I., and Marcus, D. (2013). Northwestern university schizophrenia data and software tool (nusdast). Frontiers in Neuroinformatics, 7(25).

[29] Grethe, J. S., Baru, C., Gupta, A., James, M., Ludaescher, B., Martone, M. E., Papadopoulos, P. M., Peltier, S. T., Rajasekar, A., Santini, S., Zaslavsky, I. N., and Ellisman, M. H. (2005). Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform, 112:100–109.

[30] Work package deliverables of the neuGRID4You (N4U) project. (2015). Available from: https://neugrid4you.eu/workshop [Accessed 10-July-2015].

[31] Training and Resources of the neuGRID4You (N4U) project. (2015). Available from: https://neugrid4you.eu/training-and-resources [Accessed 10-July-2015].

[32] Halfond, W., Viegas, J., and Orso, A. (2006). A classification of sql-injection attacks and countermeasures. In Proceedings of the IEEE International Symposium on Secure Software Engineering, pages 65–81.